УДК 519.2

В.Б. Бериков

КОЛЛЕКТИВ АЛГОРИТМОВ С ВЕСАМИ В КЛАСТЕРНОМ АНАЛИЗЕ РАЗНОРОДНЫХ ДАННЫХ[1]

Для кластерного анализа разнородных данных предложен метод построения коллективного решения с учетом весов различных алгоритмов. Введена вероятностная модель ансамблевого кластерного анализа с латентными классами, учитывающая веса. В рамках модели получено выражение для верхней границы ошибки классификации. Предложен способ выбора весов, для которых эта граница принимает минимальное значение. С помощью статистического моделирования продемонстрирована эффективность предложенного метода.

Ключевые слова: *кластерный анализ*, *коллективное принятие решений*, *алгоритмы с весами*, *вероятность ошибки классификации.*

Задача кластерного анализа (таксономия, группировка объектов по похожести их характеристик, автоматическая классификация «без учителя») может быть сформулирована следующим образом [1–3]. Имеется множество объектов, описываемых набором некоторых переменных (либо матрицей попарных расстояний). Эти объекты требуется разбить на относительно небольшое число кластеров (таксонов, групп, классов) так, чтобы критерий качества группировки принял бы наилучшее значение. Число кластеров может быть как выбрано заранее, так и не задано (в последнем случае оптимальное количество кластеров должно быть определено автоматически). Под критерием качества обычно понимается некоторый функционал, зависящий от разброса объектов внутри группы и расстояний между группами.

В последнее время в кластерном анализе активно развивается подход, основанный на коллективном принятии решений [4, 5]. Известно, что алгоритмы кластерного анализа не являются универсальными: каждый алгоритм имеет свою специфическую область применения: например, одни алгоритмы лучше справляются с задачами, в которых объекты каждого кластера описаны «шарообразными» областями многомерного пространства; другие алгоритмы предназначены для поиска «ленточных» кластеров и т.д. В случае, когда данные имеют разнородную природу (см. пример на рис. 1), для выделения кластеров целесообразно применять не какой-то один алгоритм, а набор различных алгоритмов. Коллективный (ансамблевый) подход позволяет также снижать зависимость результатов группировки от выбора параметров алгоритма, получать более устойчивые решения в условиях «зашумленных» данных, при наличии в них «пропусков».

Существуют следующие основные методики получения коллективных кластерных решений [5]: использование матрицы попарного сходства/различия объектов; максимизация степени согласованности решений (нормализованной взаимной информации, исправленного индекса Ранда и т.д.); применение теоретико-

[1] Работа выполнена при поддержке РФФИ, проекты 11-07-00346а, 10-01-00113а, 11-07-12083-офи-м-2011.

графовых методов; анализ бутстрэп-выборок. В предлагаемой работе развивается направление, основанное на матрицах попарного различия объектов. В итоговом коллективном решении требуется учитывать степень «компетентности» каждого алгоритма на различных подмножествах объектов. Для оценивания компетентности используется модель ансамблевого кластерного анализа, предложенная в работе [6]. Вводится модификация данной модели, учитывающая веса алгоритмов. Предложенная модель используется для обоснования качества коллективного решения.

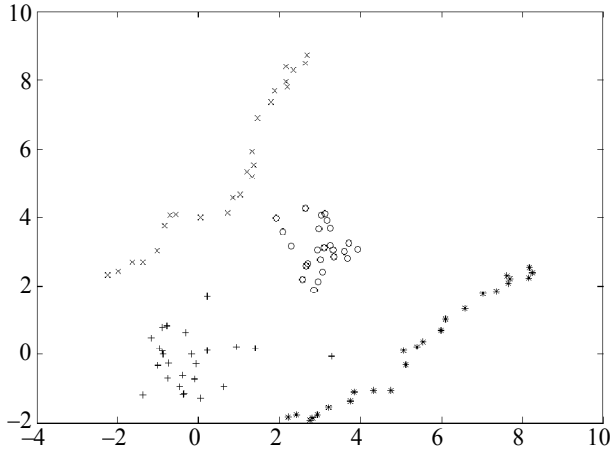

Рис. 1. Пример расположения данных

Заметим, что вопросы, связанные с теоретическим обоснованием качества группировки, остаются одними из наиболее важных в кластерном анализе.

1. Основные понятия и обозначения

Пусть имеется множество $s = \{o^{(1)},...,o^{(N)}\}$ некоторых объектов, случайным и независимым образом выбранных из генеральной совокупности. Требуется разбить эти объекты на заданное число K кластеров в соответствии с критерием качества группировки.

Пусть каждый объект описывается с помощью набора вещественных переменных $X_1,\ldots,X_n$. Через $x = x(o) = (x_1,...,x_n)$ обозначим вектор переменных для объекта o, где $x_j = X_j(o)$, $j = 1,...,n$, а через $\boldsymbol{x}_N$ – таблицу данных $(x(o^{(1)},...,x(o^{(N)}))^T$. Предположим, что имеется некоторая скрытая (непосредственно ненаблюдаемая) переменная Y, которая задает принадлежность каждого объекта к некоторому классу $Y \in \{1,...,K\}$. Каждый класс характеризуется определенным законом условного распределения $p(x \mid Y = k) = p_k(x)$, $k = 1,...,K$.

Рассмотрим следующую вероятностную модель генерации данных. Пусть для каждого объекта определяется класс, к которому он относится, в соответствии с априорными вероятностями $P_k = P(Y = k)$, $k = 1,...,K$, где $\sum_{k=1}^{K} P_k = 1$. Затем в со-

ответствии с распределением $p_k(x)$ определяется значение x. Указанная процедура проводится независимо для каждого объекта. Для произвольной пары объектов $a,b \in s$ и соответствующих наблюдений $x(a)$ и $x(b)$ определим величину

$$Z = I(Y(a) \neq Y(b)),$$

где $I(\cdot)$ – индикаторная функция ($I(true) = 1$, $I(false) = 0$). Обозначим через

$$P_Z = P[Z = 1 \mid x(a), x(b)]$$

вероятность события «a и b принадлежат к различным классам, при известных $x(a)$ и $x(b)$»:

$$P_Z = 1 - P[Y(a) = 1 \mid x(a)]P[Y(b) = 1 \mid x(b)] - ... - P[Y(a) =$$
$$= K \mid x(a)]P[Y(b) = K \mid x(b)] = 1 - \sum_{k=1}^{K} \frac{p_k(x(a))p_k(x(b))P_k^2}{p(x(a))p(x(b))},$$

где $p(x(o)) = \sum_{k=1}^{K} p_k(x(o))P_k$, $o = a,b$.

Пусть с помощью некоторого набора алгоритмов кластерного анализа $\mu_1,...,\mu_M$ по таблице данных строятся варианты разбиения множества s на кластеры (число кластеров для каждого варианта может отличаться). Поскольку нумерация классов не играет роли, удобнее рассматривать отношение эквивалентности, т.е. устанавливать, относит ли алгоритм μ_m каждую пару объектов в один и тот же класс, либо в разные классы. Определим для каждой пары объектов a и b величину $h_m(a,b) = I[\mu_m(a) \neq \mu_m(b)]$.

Рассмотрим следующую модель ансамблевого кластерного анализа. Предположим, каждый алгоритм μ_m рандомизирован, т.е. зависит от случайного вектора Ω_m, принадлежащего некоторому множеству $\boldsymbol{\Omega}_m$ (параметров, или, в более общем смысле, «условий обучения», таких, как порядок объектов, подмножество отобранных переменных или случайная подвыборка объектов и т.п.). Кроме того, будем считать, что решения алгоритма зависят от действительного статуса пары a,b (т.е. от Z), а также от исходной таблицы данных: $h_m(a,b) = h_m(a,b,\boldsymbol{x}_N,\Omega_m,Z)$. В дальнейшем, при фиксированных a, b и $\boldsymbol{x}_N$ будем обозначать $h_m(a,b,\boldsymbol{x}_N,\Omega_m,Z) = h_m(\Omega_m,Z)$. Предположим, что выполняется

$$P[h_m(\Omega_m,Z) = 1 \mid Z = 1] = P[h_m(\Omega_m,Z) = 0 \mid Z = 0] = q_m,$$

т.е. условная вероятность правильного решения (либо разделения, либо объединения пары объектов) для алгоритма μ_m постоянна. Кроме того, будем полагать, что величина $q_m > 0,5$. В литературе это предположение известно как условие достижения «слабой обученности» алгоритмов. Это означает, что каждый алгоритм дает группировку, для которой вероятность правильного решения выше, чем у тривиального алгоритма, основанного на случайном равновероятном выборе в пользу объединения или разделения рассматриваемой пары объектов. Обозначим

$$P_m = P[h_m(\Omega_m,Z) = 1].$$

Чем ближе величина P_m к 0 или 1 (или чем меньше дисперсия m-го алгоритма $V_m = P_m(1-P_m)$), тем более однородными являются решения в ансамбле для алгоритма μ_m.

Предположим, что алгоритм μ_m проработал L_m раз при независимых и одинаково распределенных «параметрах». Через $\Omega_{1,m},...,\Omega_{L_m,m}$ обозначим независимые статистические копии случайного вектора Ω_m. В результате работы получим набор случайных решений

$$h_m(\Omega_{1,m}, Z),\ ...,\ h_m(\Omega_{L_m,m}, Z),\ m = 1,...,M\ .$$

Для всех $\Omega_{l,m}$ каждый алгоритм μ_m работает независимо (не использует результаты, полученные для других $\Omega_{l',m}$, $l' \neq l$). Разные алгоритмы также независимы в том смысле, что они не используют результаты, полученные другими алгоритмами. С другой стороны, решения являются зависимыми от действительного статуса пары a,b (например, если объекты a и b принадлежат разным кластерам, «разумные» алгоритмы с большей вероятностью будут относить эти объекты также к различным классам). Данное свойство может быть формализовано следующим образом. Предположим, что решения являются условно независимыми:

$$P[h_{m_1}(\Omega_{i_1,m_1}, Z) = h_{r_1},...,h_{m_j}(\Omega_{i_j,m_j}, Z) = h_{r_j} \mid Z = z] =$$

$$= P[(h_{m_1}(\Omega_{i_1,m_1}, Z) = h_{r_1} \mid Z = z] \cdot ... \cdot P[h_{m_j}(\Omega_{i_j,m_j}, Z) = h_{r_j} \mid Z = z],$$

где $\Omega_{i_1,m_1},...,\Omega_{i_j,m_j}$ – произвольный набор параметров, а индексы $m_1,...,m_j$ соответствуют различным алгоритмам, $h_{r_1},...,h_{r_j}, z \in \{0,1\}$.

Пусть

$$P_{m,m} = P[h_m(\Omega'_m, Z) = 1, h_m(\Omega''_m, Z) = 1],$$

где Ω'_m, Ω''_m принадлежат тому же распределению, что и Ω_m, $m = 1,...,M$, $\Omega'_m \neq \Omega''_m$.

Будем полагать, что каждый алгоритм вносит свой вклад в общее коллективное решение. Обозначим

$$\overline{H} = \sum_{m=1}^{M} \alpha_m \frac{1}{L_m} \sum_{l=1}^{L} h_m(\Omega_{l,m}, Z),$$

где α_m – некоторые константы (веса), $m = 1,...,M$, $\sum_m \alpha_m = 1$. Функцию

$$c(h_1(\Omega_1, Z),...,h_m(\Omega_l, Z),...,h_M(\Omega_L, Z)) = I\,[\overline{H} > 1/2]$$

назовем ансамблевым решением для a и b, полученным в соответствии со «взвешенным голосованием». Для построения окончательной ансамблевой кластеризации объектов можно использовать различные подходы [5]. Например, можно применять метод, основанный на матрице попарных различий

$$\boldsymbol{H} = \left(\overline{h}(o^{(i_1)}, o^{(i_2)})\right),$$

где $o^{(i_1)}, o^{(i_2)} \in s$, $o^{(i_1)} \neq o^{(i_2)}$, величина $\overline{h}$ есть наблюдаемое значение $\overline{H}$. Эту матрицу можно рассматривать как матрицу попарных расстояний между объекта-

ми и применять алгоритм построения дендрограммы [2] для разбиения множества объектов на заданное число кластеров.

2. Верхняя граница вероятности ошибочной классификации

Рассмотрим *маржинальную функцию* кластерного ансамбля [6]:

$mg = \{$ взвешенное число голосов за Z $-$ взвешенное число голосов против $Z\}$,

где $Z \in \{0,1\}$. Легко показать, что маржинальная функция

$$mg = mg(\overline{H}, Z) = (2Z-1)(2\overline{H}-1).$$

С использованием этой функции, можно представить вероятность ошибки предсказания Z как

$$P_{\text{err}} = P_{Z,\Omega_{1,1},\ldots,\Omega_{L_M,M}}[mg(\overline{H}, Z) < 0].$$

Из неравенства Чебышева следует

$$P_{\text{err}} < \frac{Var\, mg(\overline{H}, Z)}{(E\, mg(\overline{H}, Z))^2}, \qquad (1)$$

при условии, что $E\, mg(\overline{H}, Z) > 0$ (через $E(\cdot)$ и $Var(\cdot)$ обозначено математическое ожидание и дисперсия соответственно).

В работе [6] для случая одного алгоритма ($M = 1$) и без учета весов были найдены выражения для математического ожидания и дисперсии маржинальной функции:

$$E\, mg(\overline{H}, Z) = 2\, q_1 - 1, \quad Var\, mg(\overline{H}, Z) = \frac{4}{L_1}(P_1 - P_{1,1}).$$

Аналогичным образом можно вывести выражения для характеристик маржинальной функции для случая нескольких алгоритмов с весами. Справедлива следующая

Теорема 1. Пусть выполняются введенные выше предположения модели. Тогда

$$E\, mg(\overline{H}, Z) = 2\sum_{m=1}^{M} \alpha_m\, q_m - 1, \quad Var\, mg(\overline{H}, Z) = 4\sum_{m=1}^{M} \frac{\alpha_m^2}{L_m}(P_m - P_{m,m}), \qquad (2)$$

причем для всех $m = 1,\ldots M$ имеет место $P_m - P_{m,m} = q_m(1-q_m)$.

Замечание. Требование $E\, mg(\overline{H}, Z) > 0$ выполняется, если $\sum_{m=1}^{M} \alpha_m\, q_m > 0{,}5$, т.е. усредненная условная вероятность правильного решения выше, чем вероятность правильного решения, полученного тривиальным алгоритмом случайного равновероятного выбора.

Данная теорема позволяет оценить сверху вероятность ошибки ансамблевого алгоритма в выражении (1), а также сделать несколько качественных выводов. Рассмотрим коэффициент корреляции ρ_m между $h'_m = h(\Omega'_m, Z)$ и $h''_m = h(\Omega''_m, Z)$, где $\Omega'_m \neq \Omega''_m$, $m = 1,\ldots,M$. По определению коэффициента

$$\rho_m = \rho_{h'_m, h''_m} = \frac{P_{m,m} - P_m^2}{P_m(1-P_m)}.$$

Так как $P_m - P_{m,m} = q_m(1-q_m)$ и $P_m - P_{m,m} = P_m - P_m^2 + P_m^2 - P_{m,m}$, то получим следующее выражение:

$$Var(mg(\overline{H}, Z)) = 4\sum_m \frac{\alpha_m^2}{L_m}(1-\rho_m)V_m \ .$$

Таким образом, можно сделать следующие выводы:

- вероятность ошибки уменьшается с ростом числа элементов L_m в ансамбле для каждого алгоритма μ_m;

- повышение однородности ансамбля (т.е уменьшение дисперсии V_m) и увеличение корреляции между его решениями снижает вероятность ошибки.

3. Оценивание характеристик ансамбля и нахождение оптимальных весов

Предположим, в результате работы ансамбля получены различные варианты классификации объектов. Для фиксированной пары объектов a и b, каждого алгоритма μ_m и l-го элемента соответствующего ансамбля определим величину $h_{m,l} \in \{0,1\}$, показывающую, была ли данная пара объектов отнесена к одному или разным кластерам, где $l = 1,...,L_m$, $m = 1,...,M$. Эти значения можно использовать для оценивания введенных выше характеристик ансамбля. Например, оценкой P_m служит величина $\tilde{P}_m = \frac{1}{L}\sum_{l=1}^{L} h_{m,l}$, а оценкой q_m – величина $\tilde{q}_m = \max(\tilde{P}_m, 1-\tilde{P}_m)$. Аналогичным образом можно оценить и другие введенные характеристики ансамбля.

Рассмотрим задачу выбора оптимальных весов $\alpha_1,...,\alpha_M$, для которых минимальна верхняя граница вероятности ошибки (1) при ограничении $\sum \alpha_m = 1$. Для ее решения можно использовать известные методы поиска глобального экстремума функций [7]. Однако поскольку оценка (1) является значительно завышенной в силу того, что при ее выводе используется неравенство Чебышева, справедливое для любого распределения, можно использовать и другие подходы. Например, минимизировать дисперсию маржинальной функции в выражении (2), при условии, что математическое ожидание не ниже заданного значения и т.п. Нахождение оптимальных коэффициентов может проводиться методом множителей Лагранжа с учетом ограничений в виде неравенств. При этом возникает задача квадратичного программирования, для которой существует достаточно разработанная теория. Более подробное изучение различных вариантов поставленной оптимизационной задачи планируется провести в дальнейших работах.

В случае, когда число алгоритмов $M = 2$, число элементов ансамбля для каждого алгоритма одинаково, а в качестве целевого минимизируемого функционала рассматривается дисперсия маржинальной функции, можно получить следующее выражение для оптимальных значений весов:

$$\alpha_1^* = \frac{\tilde{q}_2(1-\tilde{q}_2)}{\tilde{q}_1(1-\tilde{q}_1)+\tilde{q}_2(1-\tilde{q}_2)}, \quad \alpha_2^* = 1-\alpha_1^* . \qquad (3)$$

Таким образом, при назначении оптимального веса учитывается, как вели себя алгоритмы при классификации данной пары объектов при различных условиях (параметрах) работы. Тот алгоритм, который давал более устойчивые решения, получает больший вес по сравнению с менее устойчивым (для которого разброс решений выше).

Интерес представляет также случай, когда для всех пар объектов веса, соответствующие каждому алгоритму, совпадают. Таким образом, требуется оценить общую степень «компетентности» каждого алгоритма по всевозможным парам. Для этого можно использовать выражение (3), подставив в него усредненные по всем парам объектов величины $\tilde{q}_1, \tilde{q}_2$.

Разработан алгоритм построения ансамблевого решения, реализующий предложенную методику определения оптимальных весов алгоритмов. При этом для каждой пары объектов с помощью приближенного переборного алгоритма находились веса, для которых верхняя граница вероятности ошибки (3) принимает минимальное значение. После вычисления согласованной матрицы различий, для нахождения итогового варианта группировки применялся стандартный агломеративный метод построения дендрограммы, который в качестве входной информации использовал попарные расстояния между объектами [2]. При этом расстояние между группами определялось по принципу «средней связи», т.е. как среднее арифметическое попарных расстояний между объектами, входящими в группы.

Напомним также, что предполагается независимость отбора различных параметров алгоритмов, а также достижение алгоритмами «слабой» обученности. Чтобы в максимальной степени обеспечить выполнение условий теоремы 1, дополнительно проводился контроль качества решений базовых алгоритмов по индексам качества результатов кластерного анализа [3]. Так, например, если в полученной группировке число объектов, попавших в один кластер, оказывалось меньше заданного порога, такой вариант решения исключался из рассмотрения.

4. Экспериментальное исследование

При исследовании разработанной методики построения ансамбля рассматривался случай, когда в ансамбль включены два алгоритма: алгоритм k-средних и агломеративный алгоритм построения дендрограммы [2], в котором расстояние между группами определялось по принципу «ближайшего соседа». Для оценки качества использовался метод статистического моделирования: многократно генерировались выборки, соответствующие заданному распределению для каждого класса; полученный набор данных классифицировался с помощью предложенного ансамблевого алгоритма с весами; вычислялся индекс согласованности полученной классификации с истинной. Для определения степени согласованности использовался индекс Ранда, представляющий собой отношение числа пар объектов, у которых либо одинаковые, либо разные номера классов в полученной и истинной группировке, к общему числу пар различных объектов (значение индекса, близкое к 1, означает хорошую согласованность группировок). Степень согласованности усреднялась по всем выборкам. Число повторений выборок было задано равным 100.

В 30-мерном пространстве переменных моделировались кластеры, два из которых имеют шарообразную, а два – ленточную форму (см. типичный пример выборки – проекцию в пространство первых двух переменных на рис. 1). Некоторые

переменные, номера которых определялись случайно, являлись «шумовыми», т.е. все реализации по этим переменным подчинялись равномерному распределению (число шумовых переменных задавалось параметром n_0). Объем выборки для каждого из четырех классов был задан равным 25. Для построения ансамбля использовался метод случайных подпространств [5]: из общего числа переменных случайным образом выбиралось заданное число n_{ans} переменных, в пространстве которых проводилась группировка. В рассмотренном примере было выбрано значение $n_{\text{ans}} = 3$. Число элементов ансамбля для каждого алгоритма было положено равным 50. На рис. 2 показаны полученные результаты моделирования, в зависимости от числа шумовых переменных. Через R^*_{ens} обозначено значение индекса Ранда для разработанного ансамблевого алгоритма с весами. Для сравнения, приведены результаты работы каждого из алгоритмов без использования ансамбля (на рисунке обозначено: R_{10} – индекс Ранда для алгоритма *k*-средних, R_{20} – для алгоритма построения дендрограммы), а также с использованием ансамбля по каждому алгоритму отдельно ($R_{1\text{ens}}$ – для алгоритма *k*-средних, $R_{2\text{ens}}$ – для алгоритма построения дендрограммы). В последних двух случаях число элементов в ансамбле задавалось равным 100, чтобы обеспечить правомочность сравнения.

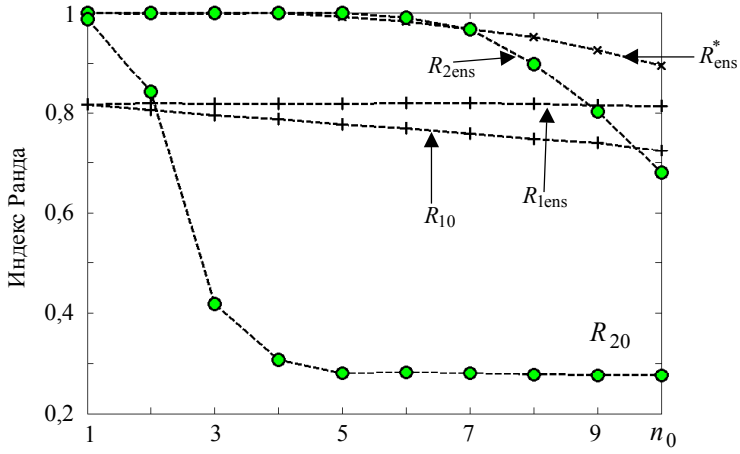

Рис. 2. Результаты моделирования

По результатам моделирования можно сделать следующие выводы. Эксперимент подтвердил преимущество коллективного подхода (особенно это проявляется для алгоритма построения дендрограммы, точность которого при использовании ансамбля возросла в несколько раз). Начиная с определенного порога на число шумовых переменных (в данном случае равного восьми), ансамблевый алгоритм с весами позволил получить наилучшую точность по сравнению с другими сравниваемыми методиками. При меньшем числе шумовых переменных точность данного алгоритма практически совпала с точностью ансамблевого алгоритма построения дендрограммы (с удвоенным числом элементов ансамбля) и также являлась оптимальной. Заметим, что заранее неизвестно, какой из двух алгоритмов

(алгоритм *k*-средних или алгоритм построения дендрограммы) оказался бы наилучшим. Использование метода взвешенного ансамбля позволило в данном примере получить гарантированный оптимальный результат.

Заключение

Одна из трудностей кластерного анализа – возможная неоднозначная интерпретация результатов группировки. Алгоритмы, в основу которых заложены различные подходы, могут дать несогласующиеся результаты. В настоящей работе предполагается, что алгоритмы, на основе которых строится коллективное решение, «неортогональны» в указанном выше смысле, но взаимно дополняют друг друга: одни компенсируют «слабые места» других. В этом случае требуется подбирать степень вклада, который вносит каждый из алгоритмов в общее решение на разных структурах данных.

Для решения этой задачи предложен метод, который при построении итогового решения учитывает поведение каждого алгоритма при кластеризации в различных условиях (или при выборе различных параметров работы). На основании этого поведения алгоритму приписывается определенный вес. Для обоснования метола введена вероятностная модель ансамблевой попарной классификации с латентными классами. Модель позволила получить оценку качества, используемую при нахождении оптимальных весов, для которых верхняя граница вероятности ошибки минимальна. На основе модели сделаны теоретические выводы о том, что при выполнении определенных разумных условий качество ансамбля улучшается с ростом числа его элементов, повышением однородности ансамбля и увеличением корреляции между его решениями.

Разработан алгоритм, в котором реализован метод построения ансамбля и вычисления оптимальных весов. Экспериментальное исследование с помощью процедуры статистического моделирования подтвердило эффективность предложенного метода: в условиях большого числа шумовых переменных точность ансамблевого алгоритма с весами оказалась в среднем выше, чем у других сравниваемых алгоритмов.

ЛИТЕРАТУРА

1. *Миркин Б.Г.* Методы кластер-анализа для поддержки принятия решений: обзор. М.: Изд. дом НИУ ВШЭ, 2011.
2. *Дуда Р.*, *Харт П.* Распознавание образов и анализ сцен. М.: Мир, 1976.
3. *Jain A.K.*, *Dubes R.C.* Algorithms for clustering data. Prentice Hall, NY, 1988.
4. *Jain A.K.* Data clustering: 50 years beyond k-means // Pattern Recognition Letters. 2010. V. 31. No. 8. P. 651–666.
5. *Ghosh J.*, *Acharya A.* Cluster ensembles // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. V. 1(4). P. 305–315.
6. *Berikov V.* A latent variable pairwise classification model of a clustering ensemble // Multiple Classifier Systems, 2011. Lecture Notes on Computer Science, LNCS 6713 / C. Sansone, J. Kittler, and F. Roli (Eds.). Springer, Heidelberg, 2011. P. 279–288.
7. *Жиглявский А.А.*, *Жилинкас А.Г.* Методы поиска глобального экстремума. М.: Наука, Физматлит, 1991.

Бериков Владимир Борисович
Институт математики им. С.Л. Соболева СО РАН
E-mail: berikov@math.nsc.ru

Поступила в редакцию 4 мая 2012 г.

Berikov Vladimir B. (Sobolev Institute of mathematics SB RAS). **Collective of algorithms with weights for clustering heterogeneous data.**

Keywords: cluster analysis, collective decision, algorithms with weights, probability of wrong classification.

The paper considers a problem of heterogeneous data clustering. Under heterogeneous data one can understand the data that contain different structures: sphere-like and strip-like clusters; various geometric figures etc. To raise the grouping quality for such types of data, we suggest using the ensemble of different clustering algorithms. When including an algorithm into the ensemble, it is assumed that the algorithm produces better results for a specific type of structures. Besides, it is supposed that the experiment is planned so that the algorithms work independently, and each algorithm is functioning on independently chosen sets of parameters (learning conditions). For the construction of final decision it is recognized the behavior of each algorithm in the ensemble, on the basis of which a weight is attributed to it. A probabilistic model of ensemble clustering with latent classes and algorithm's weights is introduced. With use of the model, an expression for the upper bound of classification error probability is derived. To minimize the bound, a method of weights selection is suggested. The procedure of ensemble construction and finding the weights is implemented in correspondent algorithm. The efficiency of the suggested method is demonstrated by making use of Monte-Carlo modeling.