ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2018 Управление, вычислительная техника и информатика № 42

УДК 519.95
DOI: 10.17223/19988605/42/5

Н.А. Игнатьев

ОБОБЩАЮЩАЯ СПОСОБНОСТЬ АЛГОРИТМОВ ПО МЕРЕ КОМПАКТНОСТИ

Рассматривается вычисление обобщающей способности семейств алгоритмов распознавания с бесконечной емкостью. Для оценки обобщающей способности предлагается использовать меру компактности, значения которой определяются в зависимости от размерности и состава набора признаков, количества удаляемых шумовых объектов и числа объектов-эталонов минимального покрытия.

Ключевые слова: мера компактности; шумовые объекты; информативные признаки; объекты-эталоны.

Обобщающая способность относится к числу основных показателей, характеризующих качество распознающих алгоритмов [1]. Эта способность проявляется в умении определять принадлежность объектов к классам, которых алгоритм не видел в процессе обучения. Проверка истинности гипотезы о компактности лежит в основе многих критериев и методов теории распознавания образов. Так, в [1] описан профиль компактности для вычисления обобщающей способности семейств алгоритмов, имеющих бесконечную емкость в пространстве VC (Вапника–Червоненкиса) [2]. Для определения принадлежности произвольного допустимого объекта к классам при использовании таких семейств алгоритмов необходимо хранить в памяти всю выборку. Представителем семейства с бесконечной емкостью является алгоритм «ближайший сосед» (БС).

В практических целях при вычислении обобщающей способности достаточно использовать локальные свойства (локальные ограничения) выборок объектов [1]. Локальным ограничением в [3] можно считать предложенный Н.Г. Загоруйко показатель компактности, определяемый по числу объектов-эталонов минимального покрытия, при котором распознавание объектов классов фиксированной выборки было корректным.

Кроме показателя компактности кандидатами для включения в набор локальных ограничений являются число шумовых объектов, размерность признакового пространства, множество объектов оболочек (подмножества граничных объектов) классов [4] по заданной метрике. Интерес представляет предельное значение размерности, при превышении которого показатель компактности [3] увеличивается. Набор признаков, определяющий предельное значение, рассматривается как информативный для используемой меры близости. Размерность выше предельной приводит к размыванию сходства между объектами выборки.

Существует потребность во введении новой меры измерения компактности с помощью безразмерных величин со значениями в [0, 1]. Значения этих величин требуются для анализа того, насколько реально получаемая (по заданной мере близости) структура обучающей выборки отличается от идеальной для распознавания. Идеальной считается структура, в которой число объектов-эталонов минимального покрытия равно числу классов.

Меру компактности можно использовать для сравнения метрик и преобразований признакового пространства по отношению «лучше» на фиксированных выборках объектов. Анализ структуры выборок основывается на использовании свойств этого отношения. Методика анализа ориентирована на количественные показатели, вычисляемые по результатам разбиения объектов классов на непересекающиеся группы [4]. Гарантией единственности разбиения по числу групп и составу входящих в них объектов служит устойчивость используемого алгоритма.

Влияние шумовых объектов на показатели обобщающей способности алгоритмов многократно рассматривалось в научных публикациях. По обширному перечню работ в [5] приводится обзор различных

методов обнаружения и удаления шумовых объектов. Большинство из этих методов ориентировано на использование правила БС.

Качество распознавания по правилу БС существенно зависит от чувствительности метрики к размерности признакового пространства. Изменение размерности связано как с отбором информативных признаков, так и с переходом к описанию объектов в пространстве из латентных признаков.

В качестве инструментария для перехода к латентным признакам в [4, 6] предлагалось использовать два типа правил иерархической агломеративной группировки исходных признаков. Первый тип ориентирован на последовательное объединение двух признаков в один путем нелинейного отображения их значений на числовую ось. Группировка по правилам второго типа производится на основе значений критерия устойчивости объектов по заданной метрике в двухклассовой задаче распознавания. По каждой группе признаков вычисляется обобщенная оценка объекта.

Методы, реализующие два типа правил иерархической группировки, можно идентифицировать как нелинейные и линейные. Нелинейные методы являются инвариантными к масштабам измерений признаков. У линейных методов свойство инвариантности отсутствует. Последовательность формирования групп и латентных признаков на их основе по двум типам правил определяет порядок по отношению степени информативности. Информативность признака вычисляется как экстремум критерия разбиения его (признака) значений на непересекающиеся интервалы в форме проверки степени истинности гипотезы: *«Множества значений признака в описании объектов из разных классов при числе интервалов, равном числу классов, не пересекаются между собой»*.

В данной работе рассматриваются непустые классы (множества) метрик, кластерные структуры обучающих выборок по которым совпадают (являются эквивалентными) по числу групп и составу входящих в них объектов. Информация о кластерной структуре позволяет вести последовательный отбор объектов-эталонов минимального покрытия, в каждом из которых определена локальная метрика. Способ вычисления весов локальных метрик аналогичен используемому в методе FRiS STOLP [3].

Для оценки обобщающей способности алгоритмов метода БС предлагается применять критерий, значения которого вычисляются в зависимости от размерности и состава набора признаков, количества удаляемых шумовых объектов и числа объектов-эталонов минимального покрытия. Оценки по критерию использовались для демонстрации устойчивости результатов отбора информативных признаков методом кросс-валидации на случайных выборках.

1. О разбиении объектов классов на непересекающиеся группы

Использование частично обученной выборки (ЧОВ) для задания условий группировки описано в [7]. Примером условия служит указание подмножества из пар объектов выборки, которые при разбиении не должны попадать в одну группу. Принадлежность объектов к непересекающимся классам служит источником дополнительной информации для исследования кластерной структуры с помощью различных мер близости.

Основные идеи приводимого ниже метода изложены в [4]. Целями разбиения объектов классов на непересекающиеся группы являются:

– вычисление и анализ значений компактности объектов классов и выборки в целом;

– поиск минимального покрытия обучающей выборки объектами-эталонами.

Рассматривается задача распознавания в стандартной постановке. Считается, что задано множество $E_0 = \{S_1, ..., S_m\}$ объектов, разделенное на l ($l > 2$) непересекающихся подмножеств (классов) $K_1, ..., K_l$, $E_0 = \bigcup_{i=1}^{l} K_i$. Описание объектов производится с помощью набора из n разнотипных признаков $X(n) = (x_1, ..., x_n)$, ξ из которых измеряются в интервальных шкалах, $(n - \xi)$ – в номинальной. На множестве объектов E_0 задана метрика $\rho(x, y)$.

Обозначим через $L(E_0, \rho)$ подмножество граничных объектов классов, определяемое на E_0 по метрике $\rho(x, y)$. Объекты $S_i, S_j \in K_t$, $t = 1, ..., l$ считаются связанными между собой ($S_i \leftrightarrow S_j$), если

$\{S \in L(E_0, \rho) \mid \rho(S, S_i) < r_i \text{ and } \rho(S, S_j) < r_j\} \neq \varnothing$, где $r_i(r_j)$ – расстояние до ближайшего от $S_i(S_j)$ объекта из CK_t ($CK_t = E_0 \backslash K_t$) по метрике $\rho(x, y)$.

Множество $G_{tv}=\{ S_{v_1},...,S_{v_c} \}$, $c \geq 2$, $G_{tv} \subset K_t$, $v < |K_t|$ представляет область (группу) со связанными объектами в классе K_t, если для любых $S_{v_i}, S_{v_j} \in G_{tv}$ существует путь $S_{v_i} \leftrightarrow S_{v_k} \leftrightarrow ... \leftrightarrow S_{v_j}$. Объект $S_i \in K_t$, $t = 1, ..., l$ принадлежит группе из одного элемента и считается несвязанным, если не существует пути $S_i \leftrightarrow S_j$ ни для одного объекта $S_j \neq S_i$ и $S_j \in K_t$. Требуется определить минимальное число непересекающихся групп из связанных и несвязанных объектов по каждому классу K_t, t = 1, ..., l.

Данная задача может рассматриваться и в альтернативной постановке (без задания признаков), если определена квадратная матрица близости $\{a_{ij}\}_{m \times m}$ между m объектами и вектор $F = (f_1, ..., f_m)$, $f_i \in \{1, ..., l\}$ принадлежности объектов к классам $K_1, ..., K_l$. Вектор F служит дополнительной информацией для задания условий группировки.

При определении минимального числа групп из связанных и несвязанных объектов классов используется $L(E_0, \rho)$ – подмножество граничных объектов (оболочка) классов по заданной метрике ρ и описание объектов в новом пространстве из бинарных признаков. Для выделения оболочки классов для каждого $S_i \in K_t$, $t = 1, ..., l$ строится упорядоченная по $\rho(x, y)$ последовательность

$$S_{i_0}, S_{i_1}, ..., S_{i_{m-1}}, S_i = S_{i_0}. \quad (1)$$

Пусть $S_{i_\beta} \in CK_t$ – ближайший к S_i объект из (1), не входящий в класс K_t. Обозначим через $O(S_i)$ окрестность радиуса $r_i = \rho(S_i, S_{i_\beta})$ с центром в S_i, включающую все объекты, для которых $\rho(S_i, S_{i_\tau}) < r_i$, $\tau = 1, ..., \beta - 1$. В $O(S_i)$ всегда существует непустое подмножество объектов

$$\Delta_i = \left\{ S_{i_\alpha} \in O(S_i) \mid \rho\left(S_{i_\beta}, S_{i_\alpha}\right) = \min_{S_{i_\tau} \in O(S_i)} \rho\left(S_{i_\beta}, S_{i_\tau}\right) \right\}. \quad (2)$$

По (2) принадлежность объектов к оболочке классов определяется как $L(E_0, \rho) = \bigcup_{i=1}^{m} \Delta_i$.

Множество объектов оболочки из $K_t \cap L(E_0, \rho)$ обозначим как $L_t(E_0, \rho) = \{S^1, ..., S^\pi\}$, $\pi \geq 1$. Значение $\pi = 1$ однозначно определяет вхождение всех объектов класса в одну группу. При $\pi \geq 2$ преобразуем описание каждого объекта $S_i \in K_t$ в $S_i = (y_{i1}, ..., y_{i\pi})$, где

$$y_{ij} = \begin{cases} 1, \rho\left(S_i, S^j\right) < r_i, \\ 0, \rho\left(S_i, S^j\right) \geq r_i. \end{cases} \quad (3)$$

Пусть по (3) получено описание объектов класса K_t в новом (бинарном) признаковом пространстве, $\Omega = K_t$, θ – число непересекающихся между собой групп объектов, $S_\mu \vee S_\eta$, $S_\mu \wedge S_\eta$ – соответственно операции дизъюнкции и конъюнкции по бинарным признакам объектов $S_\mu, S_\eta \in K_t$. Пошаговое выполнение алгоритма разбиения объектов K_t на непересекающиеся группы $G_1, ..., G_\theta$ таково.

Шаг 1: $\theta = 0$.

Шаг 2: Выделить объект $S \in \Omega$, $\theta = \theta + 1$, $Z = S$, $G_\theta = \varnothing$.

Шаг 3: **Выполнять** Выбор $S \in \Omega$ and $S \wedge Z = true$, $\Omega = \Omega \backslash S$, $G_\theta = G_\theta \cup S$, $Z = Z \vee S$ **пока** $\{S \in \Omega \mid S \wedge Z = true\} \neq \varnothing$.

Шаг 4: Если $\Omega \neq \varnothing$, то идти 2.

Шаг 5: Конец.

Разбиение объектов E_0 на непересекающиеся группы по описанному выше алгоритму используется для поиска минимального покрытия [4] обучающей выборки объектами-эталонами. Обозначим через $R_S = \rho(S, \overline{S})$ расстояние от объекта $S \in K_t$ до ближайшего объекта $\overline{S}$ из противоположного к K_t класса ($\overline{S} \in CK_t$), через δ – минимальное число непересекающихся групп из связанных и несвязанных объектов классов на E_0.

Упорядочим объекты каждой группы $G_u \cap K_t$, $u = 1, ..., \delta$, $t = 1, \ldots, l$ по множеству значений $\{R_S\}_{S \in G_u}$.

В качестве меры близости между $S \in G_u$, $u = 1, ..., \delta$ и произвольным допустимым объектом S' используется взвешенное расстояние по локальной метрике $d(S, S') = \rho(S, S')/R_S$. Решение о принадлежности S' к одному из классов $K_1, ..., K_i$ принимается по правилу: $S' \in K_t$ если

$$d(S_\mu, S') = \min_{S_j \in E_0} d(S_j, S') \text{ and } S_\mu \in K_t \text{ and } d(S_\mu, S') \neq \min_{S_j \in CK_t} d(S_j, S'). \quad (4)$$

Согласно принципа *последовательного исключения*, используемого в процессе поиска покрытия, выборка E_0 делится на два подмножества: множество эталонов E_{ed} и контрольное множество E_k, $E_0 = E_{ed} \cup E_k$. В начале процесса $E_{ed} = E_0$, $E_k = \varnothing$. Упорядочение по значениям из $\{R_S\}_{S \in G_u}$, $u = 1, ..., \delta$ используется для определения кандидата на удаление из числа объектов-эталонов по группе G_u. Идея отбора заключается в поиске минимального числа эталонов, при котором алгоритм распознавания по (4) остается корректным (без ошибок распознающим объекты) на E_0.

Будем считать, что нумерация групп объектов отражает порядок $|G_1| \geq ... \geq |G_\delta|$ и по группе G_p, $p = 1, ..., \delta$ не производился отбор объектов-эталонов. Кандидаты на удаление из E_{ed} последовательно выбираются начиная с $S \in G_p$ с минимальным значением R_S. Если включение S в E_k нарушает корректность решающего правила (4), то S возвращается в множество E_{ed}.

2. О мерах компактности в задачах распознавания с учителем

Меры компактности востребованы для оценки обобщающей способности распознающих алгоритмов. При вычислении оценок используются результаты поиска и удаления шумовых объектов, отбора информативных наборов признаков, число объектов-эталонов минимального покрытия обучающих выборок. Рассмотрим метод формирования множества шумовых объектов, мощность которого зависит от проверки предлагаемого ниже условия.

Пусть $S_k \in K_i$, $\rho(S_k, S_r) = \min_{S_j \in CK_i} \rho(S_k, S_j)$ и $Z = |\{S_\mu \in K_i | \rho(S_k, S_\mu) < \rho(S_k, S_r)\}|$. Обозначим через D_i ($D_i \in CK_i$) множество шумовых объектов класса K_i. Объект $S_r \in CK_i$ включается в D_i и рассматривается как шумовой, если выполняется условие:

$$\frac{ZZ - \lambda}{|K_i|} > \frac{1}{m - |K_i|}, \quad (5)$$

где $ZZ = |\{S_\mu \in K_i | \rho(S_r, S_k) < \rho(S_p, S_k) < \rho(S_\eta, S_k)\}|$, $|\lambda| < \min_{1 \leq i \leq l} |K_i|$, $\rho(S_\eta, S_k) = \min_{S_j \in CK_i \setminus \{S_r\}} \rho(S_j, S_k)$. Значения Z и $Z + ZZ$ можно рассматривать как число представителей класса K_i в гипершаре с центром в $S_k \in K_i$ соответственно до и после удаления шумового объекта S_r.

Селекция объектов обучающих выборок при некоторых ограничениях способствует повышению обобщающей способности алгоритмов распознавания. Считается, что обобщающая способность алгоритма повышается, если дать ему возможность ошибаться на определяемых объектах выборки. В нашем случае в качестве таковых рассматриваются объекты из $\bigcup_{i=1}^{l} D_i$.

Пусть представители класса $K_i \cap \left(E_0 \setminus \bigcup_{j=1}^{l} D_j \right)$, $i = 1, \ldots, l$ разделены на минимальное число μ непересекающихся групп объектов по алгоритму из п. 1, $m_{ij} = |G_{ij}|$, $j = 1, \ldots, \mu$, $\sum_{j=1}^{\mu} m_{ij} = m_i$. Для анализа результатов разбиения класса K_i на непересекающиеся группы с учетом их числа, представительности (по количеству объектов) и удаления шумовых объектов предлагается использовать такую структурную характеристику, как оценка компактности:

$$\Theta_i = \frac{\sum_{j=1}^{\mu} m_{ij}^2}{m_i^2}. \qquad (6)$$

Очевидно, что множество допустимых значений Θ_i по (6) лежат в интервале $\left[\frac{1}{m_i}, 1\right]$. Если группа G_{i1} содержит все объекты из $K_i \cap \left(E_0 \setminus \bigcup_{j=1}^{l} D_j\right)$, то $\Theta_i = 1$. Усредненная оценка компактности обучающей выборки в целом производится с учетом доли $\left(\frac{\left|E_0 \setminus \bigcup_{i=1}^{l} D_i\right|}{m}\right)$ исключенных из рассмотрения по (5) шумовых объектов как

$$R(E_0, \rho) = \left(\frac{\left|E_0 \setminus \bigcup_{i=1}^{l} D_i\right|}{m}\right) \frac{\sum_{i=1}^{l} m_i \Theta_i}{\left|E_0 \setminus \bigcup_{i=1}^{l} D_i\right|} = \frac{\sum_{i=1}^{l} m_i \Theta_i}{m}. \qquad (7)$$

Значения (6) и (7) косвенно свидетельствуют об однородности (неоднородности) структуры обучающей выборки. Чем ближе сходство групп по числу входящих в них объектов класса, тем ближе значение (6) к $\frac{1}{m_i}$, а (7) – к $\frac{l}{m}$.

Очевидно, что число и состав шумовых объектов зависят как от значения параметра λ в (5), так и от наборов признаков в описании объектов. Проблемой реализации вычислительных процедур является согласование процессов отбора информативных признаков и удаления шумовых объектов.

Пусть структура объектов классов на выборке E_0 вычисляется по алгоритму группировки из п. 1. Обозначим через $Sh(\lambda, X(k))$ число шумовых объектов E_0, определяемых в зависимости от значения λ по (5) на наборе признаков $X(k) \subset X(n)$, CF – число объектов-эталонов минимального покрытия обучающей выборки, из которой удалены $Sh(\lambda, X(k))$ шумовых объектов. Так как невозможно получить точное решение задачи отбора информативных признаков без перебора всех их сочетаний с учетом удаления шумовых объектов, на практике рекомендуется использовать различные эвристические методы.

Независимо от используемых методов качество отбора информативных признаков предлагается определять путем проверки двух условий:

– при удалении шумовых объектов $Sh(\lambda, X(k))$ из E_0 показатель минимального покрытия выборки объектами-эталонами

$$F\left(X(k), \lambda\right) = \left(\frac{m - Sh\left(\lambda, X(k)\right)}{m}\right)\left(\frac{m - Sh\left(\lambda, X(k)\right)}{CF}\right) \qquad (8)$$

стремится к максимальному допустимому значению $\frac{m}{l}$;

– произведение числа объектов-эталонов минимального покрытия на размерность признакового пространства

$$\frac{k \times CF}{m - Sh\left(\lambda, X(k)\right)} \to \min_{E_0}. \qquad (9)$$

Первое условие (8) необходимо для оценки компактности покрытия выборки объектами-эталонами, второе (9) – для оценки сложности вычислений.

Для поиска информативных наборов $\{X(k) | X(k) \subset X(n)\}$ предлагается два критерия. Оба критерия явно не используют число объектов-эталонов минимального покрытия *CF*. Число шумовых объектов $Sh(\lambda, X(k))$ по (5) вычисляется по фиксированному значению λ. Такое λ для всех наборов $X(k) \subset X(n)$, $k \geq 2$ определяется как

$$\lambda = \arg \max_{0 \leq |\eta| < \min_{1 \leq i \leq l} |K_i|} F\left(X\left(n\right), \eta\right). \tag{10}$$

Использование (10) основано на предположении, что вероятность отбора информативных наборов признаков с более высоким значением компактности по (8) близка к нулю при λ, отличной от (10). В первом (в порядке изложения) критерии используются результаты покрытия объектов выборки гипершарами с учетом удаления шумовых объектов, во втором – оценки компактности по (7) на основе свойства связанности по объектам оболочек классов.

Пусть $O(S_i, X(k))$ $(1 \leq k < n)$ – окрестность объекта $S_i \in E_0 \cap K_j$, $j = 1, \ldots, l$, определяемая как $O(S_i, X(k)) = \{S \in K_j | \rho(S, \overline{S_i}) < \rho(S_i, \overline{S_i})\}$, где $\overline{S_i} \in CK_j$ – ближайший к S_i объект по метрике $\rho(x, y)$ из дополнения к классу K_j по множеству признаков $X(k)$. Определим оценку $S_i \in E_0$ на $X(k)$ как

$$Z(S_i, X(k)) = \max_{S_i \in O(S, X(k))} |O(S, X(k))|. \tag{11}$$

Признак $x_d \in X(n)$ является кандидатом на включение в набор $X(k)$, если

$$\sum_{S_i \in T} Z\left(S_i, X\left(k+1\right)\right) > \sum_{S_i \in T} Z\left(S_i, X\left(k\right)\right), \tag{12}$$

где $X(k+1) = X(k) \cup \{x_d\}$, $T \subset E_0$.

Обозначим через P подмножество индексов признаков из $X(n)$; $D_j(P)$ – множество шумовых объектов класса K_j по (5) на наборе $\{x_a\}_{a \in P}$ при значении λ, вычисленное по (10). Пошаговый отбор информативных наборов признаков с использованием (11) и (12) реализуется следующим образом.

Шаг 1: Выбор $i1, j1 \in \{1, ..., n\}$. $P = \{i1, j1\}$.

Шаг 2: Выделить $\bigcup_{j=1}^{l} D_j\left(P\right)$ по (5) на $\{x_a\}_{a \in P}$. $T = E_0 \backslash \bigcup_{j=1}^{l} D_j\left(P\right)$. Вычислить $\theta(P) = \left\{\theta_i\left(P\right)\right\}_1^m$ по $\{x_a\}_{a \in P}$,

где $\theta_i(P) = \{S_\mu, S_i \in K_j | \rho(S_i, S_\mu) < r_i, r_i = \min_{S_t \in CK_j \cap T} \rho(S_i, S_t)\}$.

Шаг 3: $u = 0$. $Z(P) = \left\{z_i\left(P\right)\right\}_1^m$, где $z_i(P) = \max_{S_i \in \theta_j(P)} |\theta_j(P)|$. $Y = 0$.

Для всех $v \in \{1, ..., n\} \backslash P$

выделить $\bigcup_{j=1}^{l} D_j\left(P \cup \{v\}\right)$ по (5) на $\{x_a\}_{a \in P \cup \{v\}}$, $T = E_0 \backslash \bigcup_{j=1}^{l} D_j\left(P \cup \{v\}\right)$, $C = \sum_{S_i \in T} z_i\left(P\right)$,

вычислить $\theta\left(P \cup \{v\}\right) = \left\{\theta_i\left(P \cup \{v\}\right)\right\}_1^m$ по $\{x_a\}_{a \in P \cup \{v\}}$, где $\theta_i\left(P \cup \{v\}\right) = \{S_\mu, S_i \in K_j | \rho(S_i, S_\mu) < r_i,$ $r_i = \min_{S_t \in CK_j \cap T} \rho(S_i, S_t)\}$;

вычислить $Z(P \cup \{v\}) = \left\{z_i\left(P \cup \{v\}\right)\right\}_1^m$, где $z_i\left(P \cup \{v\}\right) = \max_{S_i \in \theta_j\left(P \cup \{v\}\right)} \left|\theta_j\left(P \cup \{v\}\right)\right|$, $N = \sum_{S_i \in T} z_i\left(P \cup \{v\}\right)$.

Если $N > C$ и $N > Y$, то $Y = N, u = v$;

Шаг 4: Если $Y > 0$, то $P = P \cup \{u\}$, идти 2.

Шаг 5: Вывод P.

Шаг 6: Конец.

Для отбора информативных наборов признаков по (7) предлагается следующий алгоритм.

Шаг 1: Выбор $i_1, j_1 \in \{1, ..., n\}$. $P = \{i_1, j_1\}$.

Шаг 2: Выделить $\bigcup_{j=1}^{l} D_j(P)$ по (5) на $\{x_a\}_{a\in P}$. $T = E_0 \backslash \bigcup_{j=1}^{l} D_j(P)$. Вычислить $O(P) = \{O_i(P)\}_{S_i \in T}$ по $\{x_a\}_{a\in P}$, где $O_i(P) = \left\{ S_\mu, S_i \in K_j \middle| \rho(S_i, S_\mu) < r_i, r_i = \min_{S_t \in CK_j \cap T} \rho(S_i, S_t) \right\}$. Вычислить разбиение на группы G_{11}, ..., $G_{l\eta}$, $\eta \geq l$ по (2) и (3) алгоритмом из п. 1, $m_{ij} = |G_{ij}|$, $m_i = \sum_j m_{ij}$, $\sum_{i=1}^{l} m_i = |T|$. Вычислить $\{\Theta_i\}_1^l$ по (6) и $C = \frac{\sum_{i=1}^{l} m_i \Theta_i}{m}$.

Шаг 3: $u = 0$. $Y = C$.

Для всех $v \in \{1, ..., n\} \backslash P$

выделить $\bigcup_{j=1}^{l} D_j(P \cup \{v\})$ по (5) на $\{x_a\}_{a\in P\cup\{v\}}$, $T = E_0 \backslash \bigcup_{j=1}^{l} D_j(P \cup \{v\})$.

Вычислить $O(P \cup \{v\}) = \{O_i(P \cup \{v\})\}_{S_i \in T}$ по $\{x_a\}_{a\in P\cup\{v\}}$, где

$$O_i(P \cup \{v\}) = \left\{ S_\mu, S_i \in K_j \middle| \rho(S_i, S_\mu) < r_i, r_i = \min_{S_t \in CK_j \cap T} \rho(S_i, S_t) \right\}.$$

Вычислить разбиение на группы G_{11}, ..., $G_{l\eta}$, $\eta \geq l$ по (2) и (3) алгоритмом из п. 1, $m_{ij} = |G_{ij}|$, $m_i = \sum_j m_{ij}$, $\sum_{i=1}^{l} m_i = |T|$. Вычислить $\{\Theta_i\}_1^l$ по (6) и $N = \frac{\sum_{i=1}^{l} m_i \Theta_i}{m}$.

Если $N > Y$, то $Y = N$, $u = v$.

Шаг 4: Если $u > 0$, то $P = P \cup \{u\}$, идти 2.

Шаг 5: Вывод P.

Шаг 6: Конец.

Для удобства дальнейшего изложения алгоритмы отбора информативных признаков (в порядке их описания) будем идентифицировать как *ALG*1 и *ALG*2. Для ослабления зависимости результатов отбора от выбора начальных приближений можно использовать модификацию этих алгоритмов. Модификация заключается в сочетании принципов пошагового включения в набор информативных признаков и удаления из набора малоинформативных признаков. Для сравнения информативных наборов, полученных по разным критериям, рекомендуется использовать (8) и (9).

3. О единственности выбора кластерной структуры на обучающей выборке

Исследование единственности выражается в доказательстве существования множеств (классов) метрик, кластерные структуры фиксированных обучающих выборок при использовании которых совпадают по числу и составу групп объектов. Утверждается, что такому требованию удовлетворяют классы эквивалентных метрик $\{\Psi\}$. Например, эквивалентной к метрике ρ_1 является метрика $\rho_2 = \frac{\rho_1}{1+\rho_1}$.

Из ρ_1, $\rho_2 \in \Psi$ следует, что отношения близости между объектами на E_0 по метрике ρ_1 остаются таковыми и по метрике ρ_2. Другим следствием эквивалентности является сходство объектов оболочек классов, ближайших объектов из противоположных классов, числа групп и их состава на E_0 при реализации алгоритма группировки из п. 1. Для вычисления меры компактности по (6) и (7) можно использовать любую метрику из класса эквивалентности Ψ.

Кластерные структуры, получаемые по разным метрикам из класса Ψ, отличаются между собой лишь конфигурацией таксонов. Конфигурация таксонов влияет на значения весов локальных метрик, используемых в (4), а следовательно, на количество и состав объектов-эталонов минимального покрытия.

Для сравнения кластерной структуры по метрике $\rho \in \Psi$ на данных, отличающихся количеством представителей классов и выборок в целом, предлагается использовать вычисление оценки по (6). Тогда компактность по обучающей выборке E_0 по набору признаков $X(k)$, $k \le n$ с учетом удаления шумовых объектов по (5) будет выглядеть так:

$$U\left(E_0, X\left(k\right), \rho, \lambda\right) = \frac{\sum\limits_{i=1}^{l}\left(1 - \frac{\left|K_i\right|}{m}\right)\left|K_i\right|\Theta_i}{l-1}. \tag{13}$$

Интерес представляет анализ результатов алгоритмов *ALG*1 и *ALG*2 при отборе информативных наборов признаков. В общем случае наборы, полученные по *ALG*1 и *ALG*2, по $\rho \in \Psi$ при совпадении номеров i_1, j_1 на первом шаге различаются друг от друга. Так как при вычислении оценок по (7) и (11) учитывается порядок следования объектов, то $\rho \in \Psi$ и множество наборов совпадают по каждому алгоритму (*ALG*1 или *ALG*2), но не между алгоритмами.

Локальные метрики объекта $S \in E_0$, формируемые из класса Ψ и используемые в (4), в общем случае не являются эквивалентными. Эта особенность класса Ψ объясняет различие числа объектов-эталонов минимального покрытия обучающей выборки и его состава.

4. Вычислительный эксперимент

Для демонстрации методики вычисления мер компактности и отбора информативных наборов признаков использовалась выборка данных GERMAN из [8]. Выборка представлена двумя непересекающимися классами K_1 (700 объектов) и K_2 (300 объектов). Объекты описываются 7 количественными и 13 номинальными признаками из набора $X(20) = (x_1, ..., x_{20})$. Для унификации масштабов измерений данных множество значений каждого количественного признаков пронормировано в [0, 1].

Зафиксируем одну метрику из класса эквивалентности Ψ и будем считать ее базовой для вычислительного эксперимента. При вычислении меры близости между объектами в качестве базовой использовалась метрика Журавлева

$$\rho(x, y) = \sum_{i \in I}\left|x_i - y_i\right| + \sum_{i \in J}\begin{cases}1, x_i \ne y_i,\\ 0, x_i = y_i,\end{cases} \tag{14}$$

где $I, J \subset \{1, ..., 20\}$ – множества номеров соответственно количественных и номинальных признаков.

Из-за особенностей вычисления расстояний по локальным метрикам объектов E_0 число объектов-эталонов минимального покрытия выборки для $\rho_1, \rho_2 \in \Psi$ в общем случае различаются. Сходство топологических структур эквивалентных метрик выражается в совпадении как числа, так и состава шумовых объектов, определяемых по (5). Связь процесса выбора параметра λ в (5) с оценками компактности (8) показана в табл. 1. В скобках указано число объектов-эталонов, вычисляемых по взвешенным расстояниям на основе эквивалентной (14) метрике $\rho^*(x, y) = \frac{\rho(x, y)}{1 + \rho(x, y)}$.

Таблица 1

Оценки компактности по метрике (14) с учетом удаления шумовых объектов

λ	Число		Оценка компактности по (8) на $X(20)$
	шумовых объектов	объектов-эталонов	
2	119	203 (200)	3,8235
1	148	171 (172)	4,2451
0	217	161 (162)	3,8080
–1	239	141 (140)	4,1072
–2	261	122 (122)	4,4764
–3	261	122 (122)	4,4764

Результаты анализа структуры выборки из табл. 1 показывают, что оптимальное отношение между числом объектов-эталонов минимального покрытия и числом удаляемых шумовых объектов по (5) на $X(20)$ достигается при значении параметра $\lambda = -2$. На всех последующих этапах эксперимента число шумовых объектов по умолчанию определяется по $\lambda = -2$ в (5).

Рассмотрим зависимость числа и состава наборов информативных признаков от выбора начальных приближений в алгоритмах *ALG*1 и *ALG*2. Каждое начальное приближение (табл. 2, табл. 3) задано парой индексов-признаков.

Таблица 2

Отбор информативных признаков алгоритмом *ALG*1

Начальное приближение	Информативный набор
$i_1 = 1, j_1 = 2$	$x_1, x_2, x_3, x_4, x_5, x_{13}, x_{14}$
$i_1 = 3, j_1 = 4$	$x_1, x_2, x_3, x_4, x_5, x_{13}, x_{14}$
$i_1 = 6, j_1 = 13$	$x_1, x_2, x_3, x_5, x_6, x_8, x_{13}, x_{14}, x_{18}, x_{20}$

Таблица 3

Отбор информативных признаков алгоритмом *ALG*2

Начальное приближение	Информативный набор	Компактность по (7)
$i_1 = 1, j_1 = 2$	$x_1, x_2, x_4, x_5, x_6, x_7, x_{12}, x_{20}$	0,6688
$i_1 = 3, j_1 = 4$	$x_1, x_3, x_4, x_5, x_6, x_9, x_{13}$	0,6947
$i_1 = 6, j_1 = 13$	$x_2, x_5, x_6, x_{13}, x_{18}$	0,6417

Анализ содержимого табл. 2 и табл. 3 показывает, что наборы признаков, полученные по алгоритмам *ALG*1 и *ALG*2, различаются при выборе одинаковых начальных приближений.

Для демонстрации методики точности алгоритмов распознавания на выборке из 1 000 объектов будем использовать набор признаков, полученный по модифицированному алгоритму *ALG*1. Смысл модификации сводится к последовательному включению в набор двух информативных и удалению одного малоинформативного признака. При выборе в качестве начального приближения пары признаков (x_{17}, x_{18}) информативный набор был представлен $X(7) = (x_1, x_2, x_3, x_4, x_5, x_{13}, x_{14})$. Значения показателей распознавания объектов по исходному $X(20)$ и информативному $X(7)$ наборам признаков с использованием базовой метрики (14) и эквивалентной ей метрике $\rho^*(x,y) = \dfrac{\rho(x,y)}{1+\rho(x,y)}$ приводятся в табл. 4.

Таблица 4

Точность распознавания по выборке GERMAN

Вычисляемые показатели	Исходный набор $X(20)$	Информативный набор $X(7)$ по метрике	
		базовой (14)	эквивалентной (14)
Число шумовых объектов	261	220	220
Число эталонов	122	109	109
Среднее (8) по эталону	4,4764	5,5816	5,5816
Число ошибок (точность, %)	156 (84,4%)	147 (85,3%)	143 (85,7%)

Совокупный эффект от использования информативных наборов признаков (см. табл.4) с учетом удаления шумовых объектов более всего заметен по значениям (8) среднего числа объектов, притягиваемых одним объектом-эталоном минимального покрытия.

Для исследования обобщающей способности алгоритмов использовалось случайное деление выборки на обучение и контроль в соотношении 9 : 1. Предварительный анализ результатов показывает, что определенное преимущество в смысле значений показателей обобщающей способности имеют наборы, полученные по алгоритму *ALG*2. Из табл. 5 видна прямая корреляционная зависимость между точностью распознавания и средним числом объектов, притягиваемых одним эталоном минимального покрытия.

Таблица 5

Обобщающая способность алгоритма по базовой метрике (14)

Набор признаков	Точность распознавания %	Среднее по эталону
Исходный $X(20)$	70,9	4,4407
$x_1, x_2, x_3, x_4, x_5, x_{13}, x_{14}$	72,46	5,7088
$x_1, x_2, x_4, x_5, x_6, x_7, x_{12}, x_{20}$	72,82	5,8802
$x_1, x_3, x_4, x_5, x_6, x_9, x_{13}$	73,27	6,0136

Оценки компактности (13) для ряда подмножеств объектов GERMAN приводятся в табл. 6. Вычисление оценок производится на исходном $X(20)$ и информативном $X(7) = (x_1, x_2, x_3, x_4, x_5, x_{13}, x_{14})$ наборах признаков.

Таблица 6

Оценки компактности по (13)

№	$\|K_1\| + \|K_2\|$	Набор признаков	
		$X(20)$	$X(7)$
1	624 + 276	0,2710	0,3183
2	633 + 267	0,2658	0,2470
3	629 + 271	0,2596	0,2739
4	641 + 259	0,2597	0,2719

Отсутствие прямой коррелированности оценок (13) между наборами $X(20)$ и $X(7)$ (см. табл. 6) объясняется тем, что существует подмножество объектов, на котором набор $X(7)$ не является информативным.

Заключение

Показаны пути повышения обобщающей способности алгоритмов распознавания через удаление шумовых объектов и отбор информативных наборов признаков с использованием критериев компактности обучающей выборки. Предложенная технология может применятся при интеллектуальном анализе данных для построения информационных моделей с использованием алгоритмов распознавания.

ЛИТЕРАТУРА

1. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. 2004. № 13. С. 5–34.
2. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. М. : Наука, 1979.
3. Загоруйко Н.Г., Кутненко О.А., Зырянов А.О., Леванов Д.А. Обучение распознаванию образов без переобучения // Машинное обучение и анализ данных. 2014. Т. 1, № 7. С. 891–901.
4. Игнатьев Н.А. Кластерный анализ данных и выбор объектов-эталонов в задачах распознавания с учителем // Вычислительные технологии. 2015. Т. 20, № 6. С. 34–43.
5. Борисова И.А., Кутненко О.А. Цензурирование ошибочно классифицированных объектов выборки // Математические методы распознавания образов – 2015 : 17-я Всерос. конф., 19–25 сент. 2015. Светлогорск, 2015.
6. Мадрахимов Ш.Ф., Саидов Д.Ю. Устойчивость объектов классов и группировка признаков // Проблемы вычислительной и прикладной математики. 2016. № 3 (5). С. 50–55.
7. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. М. : Финансы и статистика, 1989. 608 с.
8. Asuncion A., Newman D.J. UCI Machine Learning Repository // University of California. Irvine. 2007. www.ics.uci.edu/mlearn/MLRepository.html.

Игнатьев Николай Александрович, д-р физ.-мат. наук, профессор. E-mail: ignatev@rambler.ru
Национальный университет Узбекистана (г. Ташкент)

Поступила в редакцию 25 июля 2017 г.

Ignatiev Nikolay A. (National University of Uzbekistan. Republic of Uzbekistan).

The generalizing ability of algorithms by the measure of compactness.

Keywords: measure of compactness; noise objects; informative features; objects-standards.

DOI: 10.17223/19988605/42/5

To estimate the generalizing ability of recognition algorithms, it is offered to use a measure of compactness. It is assumed that a training sample $E_0 = \{S_1,...,S_m\}$ is defined, divided by disjoint classes $K_1, ..., K_l$, $l \geq 2$. The objects of E_0 are described by a set of different-type features of $X(n) = (x_1, ..., x_n)$. The compactness value depends on the dimension and composition of the feature set, the number of noise objects to be deleted, and the number of objects-standards of the minimal coverage of E_0.

The compactness measure on the sample E_0 in the set of features $X(k) \subset X(n)$ $(k \leq n)$ is calculated as

$$F\left(X(k),\lambda\right) = \left(\frac{m - Sh\left(\lambda, X(k)\right)}{m}\right)\left(\frac{m - Sh\left(\lambda, X(k)\right)}{CF}\right),$$

where CF is the number of objects-standards of the minimal coverage of the sample in which $Sh(\lambda, X(k))$ noise objects are removed.

Let $S_k \in K_i$, $\rho(S_k, S_r) = \min\limits_{S_j \in CK_i} \rho(S_k, S_j)$ and $Z = |\{S_\mu \in K_i \,|\, \rho(S_k, S_\mu) < \rho(S_k, S_r)\}|$ is the number of objects in the hypersphere with the center in S_k. The object $S_r \in CK_i$ is considered as the noise object if the condition holds

$$\frac{ZZ - \lambda}{|K_i|} > \frac{1}{m - |K_i|},$$

where $ZZ = |\{S_\mu \in K_i |\ \rho(S_r, S_k) < \rho(S_p, S_k) < \rho(S_\eta, S_k)\}|$, $|\lambda| < \min\limits_{1 \leq i \leq l} |K_i|$, $\rho(S_\eta, S_k) = \min\limits_{S_j \in CK_i \setminus \{S_r\}} \rho(S_j,\ S_k)$. The ZZ value is the number of representatives of the class K_i added to the hypersphere with center at $S_k \in K_i$ after removing the noise object S_r.

To find informative sets $\{X(k) \mid X(k) \subset X(n)\}$, two criteria are proposed. Both criteria do not explicitly use the number of objects-standards of minimum coverage CF. The generalizing ability of algorithms was calculated by the method of Cross Validation on the initial and informative sets of features. The highest values were on the sets obtained according to the criterion

$$R\left(E_0, \rho\right) = \frac{\sum\limits_{i=1}^{l} m_i \Theta_i}{m} \rightarrow \max,$$

where m_i is the number of K_i objects after removing the noise objects, Θ_i is the compactness which calculated by the minimal number of disjoint groups of objects of class K_i by the metric ρ. The set of admissible values $R(E_0, \rho)$ belongs to (0, 1] and can be interpreted in terms of fuzzy logic.

A direct correlation is shown between values by the method of Cross Validation and the average number of objects attracted by the target object of the minimum coverage of the training sample. It is concluded that a measure of compactness $F(X(k), \lambda)$ can serve as an indicator of the generalizing ability. This measure is recommended for evaluating the quality of recognition algorithms in the data mining.

REFERENCES

1. Vorontsov, K.V. (2004) Kombinatornyy podkhod k otsenke kachestva obuchaemykh algoritmov [A combinatorial approach to assessing the quality of training algorithm]. In: Lupanov, O.B. (ed.) *Matematicheskie voprosy kibernetiki* [Mathematical questions of cybernetics]. Vol. 13. pp. 5–36.
2. Vapnik, V.N. (1979) *Vosstanovlenie zavisimostey po empiricheskim dannym* [Restoration of dependencies on empirical data]. Moscow: Nauka, 448 p. (In Russian).
3. Zagoruiko, N.G., Kutnenko, O.A., Zyryanov, A.O. & Levanov, D.A. (2014) Learning to recognition without overfitting. *Mashinnoe obuchenie i analiz dannykh*. 1(7). pp. 891–901. (In Russian).
4. Ignatiev, N.A. (2015) Cluster analysis and choice of standard objects in supervised pattern recognition problems. *Vychislitel'nye tekhnologii – Computational Technologies*. 20(6). pp. 34–43. (In Russian).
5. Borisova, I.A. & Kutnenko, O.A. (2015) [Censoring of erroneously classified sample objects]. *Matematicheskie metody raspoznavaniya obrazov* [Mathematical Methods Of Patterns Recognition]. The 17th All -Russian Conference. Svetlogorsk. September 19–25, 2015. (In Russian).
6. Madrakhimov, Sh.F. & Saidov, D.Y. (2016) Stability of object classes and grouping features. *Problemy vychislitel'noy i prikladnoy matematiki – Problems of Computational and Applied Mathematics.* 3(5). pp. 50–55. (In Russian).
7. Ayvazyan, S.A., Buchstaber, V.M., Yenyukov, I.S. & Meshalkin, L.D. (1989) *Prikladnaya statistika. Klassifikatsiya i snizhenie razmernostiApplied statistics* [Classification and reduction of dimensionality]. Moscow: Finansy i statistika.
8. Asuncion, A. & Newman, D.J. **(**2007**)** *UCI Machine Learning Repository*. Irvine: University of California.