

Н. А. Игнатьев, д-р физ.-мат. наук, проф., e-mail: n_ignatev@rambler.ru,
Национальный университет Узбекистана им. М. Улугбека

Компактность объектов классов и селекция обучающих выборок

Рассматривается вычисление оценок компактности обучающей выборки и минимальное покрытие ее объектами-эталоны. Для селекции выборок разработан способ удаления шумовых объектов и алгоритм отбора информативных наборов признаков с учетом числа таких объектов. При сравнении различных наборов признаков предложено использовать среднее число объектов выборки без шумовых объектов, притягиваемых одним эталоном минимального покрытия.

Ключевые слова: шумовые объекты, информативные признаки, селекция обучающей выборки, объекты-эталон, обобщающая способность, скользящий контроль

Введение

Гипотеза о компактности является центральным понятием в теории распознавания образов, и единого подхода [1] для оценки компактности не существует. Предложенная в работе [2] мера компактности отдельного класса вычисляется через геометрическую близость входящих в него объектов. Эта мера рассчитана на описание объектов количественными признаками и сильно зависит от масштабов их измерений.

Для вычисления значений компактности необходимо задание меры расстояния между объектами. В качестве такой меры чаще всего выбираются функции со свойствами метрики. Метрики определяют отношения между объектами, которые могут сильно различаться в зависимости от используемых наборов признаков в их описании.

Согласно работе [1] задача выбора информативных признаков состоит в поиске подпространства, в котором характеристика компактности достигает максимального значения. Утверждается, что при выборе набора (подмножества) признаков компактность является критерием информативности более мощным, чем распространенный минимум ошибок при кросс-валидации (скользящем контроле). Для проверки гипотезы о локальной компактности было предложено использовать функцию конкурентного сходства (FRiS-функцию). Число опорных точек или "прецедентов", необходимых для безошибочного распознавания обучающей последовательности, рассматривалось в качестве критерия информативности подсистемы признаков.

Практический интерес характеристики компактности представляют для реализации про-

цесса селекции обучающих выборок [3] в целях повышения обобщающей способности распознающих алгоритмов. Несколько способов вычисления компактности по объектам классов, столпам (объектам-эталонам) выборки, k ближайшим соседям с использованием FRiS-функций описаны в работе [4]. Там же приводится краткий обзор критериев выделения шумовых объектов из выборки. Отмечается, что большая часть этих критериев ориентирована на использование правила ближайшего соседа.

Компактность ассоциируется с непересекающимися областями признакового пространства, каждая из которых представлена группой объектов из одного класса. Так, в методе, описанном в работе [5], конфигурация областей не имеет заранее определенной формы, как правило, гиперсферы или гиперпараллелепипеда. Число групп и их состав не зависят от выбора порядка предъявления объектов для вычисления. Это свойство (устойчивости алгоритма) является следствием использования отношения связанности объектов группы через оболочку (подмножество граничных объектов) класса.

Предобработка данных, реализуемая алгоритмом группировки, гарантирует единственность числа объектов-эталонов минимального покрытия обучающей выборки [5]. Отбор объектов-эталонов проводится по каждой группе отдельно, начиная с самой представительной. Последовательность отбора в группе определяется по упорядоченным значениям отступов до ближайших объектов из противоположных классов.

Цензурирование обучающей выборки путем последовательного удаления шумовых объектов алгоритмом FRiS-Compactor описано в работе [3]. По эвристическим соображениям

устанавливалась доля таких объектов в общей выборке. Доказательство существования зависимости между числом удаляемых шумовых объектов и обобщающей способностью алгоритмов распознавания не приводится.

Наличие ряда закономерностей в данных позволяет уменьшать комбинаторную сложность алгоритмов отбора информативных признаков. Уменьшение сложности происходит за счет исключения просмотра вариантов, не ведущих к оптимальному (согласно выбранному критерию) решению. Для достижения этой цели в работе [6] было предложено использовать предобработку данных путем формирования последовательности признаков, упорядоченных по степени их независимости. Экспериментально было доказано, что значение произведения числа объектов-эталонов локально-оптимального покрытия обучающей выборки на размерность признакового пространства монотонно не возрастает при последовательном удалении малоинформативных признаков из упорядоченной последовательности.

Идея использования меры компактности обучающей выборки со множеством значений в $[a; 1]$, $0 < a < 1$, для отбора информативных латентных признаков описана в работе [7]. Пространство из латентных признаков формируется за счет нелинейного отображения значений исходных признаков в описании объектов на числовую ось с помощью правил иерархической агломеративной группировки. Последовательность объединения исходных признаков на каждом шаге группировки и порядок формирования набора латентных признаков определяются методом динамического программирования. В форме вычислительного эксперимента доказано, что значение меры компактности по первым двум латентным признакам выше, чем на исходном признаковом пространстве. Влияние способа выбора и удаления шумовых объектов на показатели обобщающей способности распознающих алгоритмов и компактности не рассматривалось.

Задача отбора информативных признаков является NP-полной. По этой причине на практике предлагается использовать различные эвристические методы. Типичным недостатком описанных в [8] методов отбора признаков является предположение о том, что признаки независимы.

Реализация многих алгоритмов машинного обучения резко усложняется, когда размерность данных велика. В работе [9] была дана геометрическая интерпретация возникнове-

ния эффекта проклятия размерности. Проклятие размерности возникает из-за того, что число возможных наборов признаков гораздо больше, чем число объектов выборки.

Целью данной работы является исследование повышения обобщающей способности алгоритмов распознавания за счет селекции обучающих выборок и снижения размерности признакового пространства. Селекцию обучающих выборок предлагается проводить путем отбора объектов-эталонов минимального покрытия и информативных наборов признаков с учетом удаления шумовых объектов.

Объектом исследования является жадный алгоритм формирования локально-оптимальных наборов признаков. Состав наборов зависит от выбора начального приближения алгоритма. Сравнение точности распознавания проводится по объектам-эталонам минимального покрытия на исходном и информативном наборах признаков.

Предметом исследования является зависимость между точностью распознавания и числом объектов выборки, притягиваемых в среднем одним объектом-эталонам минимального покрытия.

1. О разбиении объектов классов на непересекающиеся группы

Основные идеи приводимого ниже метода изложены в работе [5]. 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 $E_0 = \{S_1, \dots, S_m\}$ объектов, разделенное на l ($l > 2$) непересекающихся подмножеств (классов) $K_1, \dots, K_l$, $E_0 = \bigcup_{i=1}^l K_i$. Описание

объектов проводится с помощью набора из n разнотипных признаков $X(n) = (x_1, \dots, x_n)$, ξ из которых измеряются в интервальных шкалах, $(n - \xi)$ — в номинальной шкале. На множестве объектов E_0 задана метрика $\rho(x, y)$.

Обозначим $L(E_0, \rho)$ — подмножество граничных объектов классов, определяемое на E_0 по метрике $\rho(x, y)$. Объекты $S_i, S_j \in K_t$, $t = 1, \dots, l$, считаются связанными между собой ($S_i \leftrightarrow S_j$), если $\{S \in L(E_0, \rho) | \rho(S, S_i) < r_i \text{ и } \rho(S, S_j) < r_j\} \neq \emptyset$,

где $r_i(r_j)$ — расстояние до ближайшего от $S_i(S_j)$ объекта из CK_t ($CK_t = E_0 \setminus K_t$) по метрике $\rho(x, y)$.

Множество $G_{iv} = \{S_{v_1}, \dots, S_{v_c}\}$, $c \geq 2$, $G_{iv} \subset K_r$, $v < |K_r|$ представляет область (группу) со связанными объектами в классе K_r , если для любых S_{v_i} , $S_{v_j} \in G_{iv}$ существует путь $S_{v_i} \leftrightarrow S_{v_k} \leftrightarrow \dots \leftrightarrow S_{v_j}$. Объект $S_i \in K_r$, $t = 1, \dots, l$, 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 $S_j \in K_r$. Требуется определить минимальное число непересекающихся групп из связанных и несвязанных объектов по каждому классу K_r , $t = 1, \dots, l$.

Данная задача может рассматриваться и в альтернативной постановке (без задания признаков), если определена квадратная матрица расстояний $\{a_{ij}\}_{m \times m}$ между m объектами и вектор $F = (f_1, \dots, f_m)$, $f_i \in \{1, \dots, l\}$ принадлежности объектов к классам $K_1, \dots, K_l$. Вектор F служит дополнительной информацией для задания условий группировки. Пример использования дополнительной информации в виде частично обученной выборки (ЧОВ) имеется в работе [10]. В выборке указывалось подмножество пар объектов, которые при разбиении не должны попадать в одну группу.

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 $S_i \in K_r$, $t = 1, \dots, l$, строится упорядоченная по $\rho(x, y)$ последовательность

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

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

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

По множеству (2) принадлежность объектов к оболочке классов определяется как

$$L(E_0, \rho) = \bigcup_{i=1}^m \Delta_i.$$

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

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

Пусть по соотношению (3) получено описание объектов класса K_r в новом (бинарном) признаковом пространстве, $\Omega = K_r$, 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 $S_\mu, S_\eta \in K_r$. Пошаговое выполнение алгоритма разбиения объектов K_r на непересекающиеся группы $G_1, \dots, G_\theta$ таково:

Шаг 1. $\theta = 0$.

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

Шаг 3. Выполнять Выбор $S \in \Omega$ и $S \wedge Z = \text{true}$,
 $\Omega = \Omega \setminus S$, $G_\theta = G_\theta \cup S$, $Z = Z \vee S$
пока $\{S \in \Omega \mid S \wedge Z = \text{true}\} \neq \emptyset$.

Шаг 4. Если $\Omega \neq \emptyset$, то переход к шагу 2.

Шаг 5. Конец.

2. О селекции обучающей выборки и ее компактности

Одним из способов повышения обобщающей способности распознающих алгоритмов является селекция обучающих выборок через поиск и удаление шумовых объектов. В данной работе множество шумовых объектов рассматривается как подмножество граничных объектов классов по заданной метрике. Множество граничных объектов $B \subset E_0$ определяется как

$$B = \left\{ S \in E_0 \mid \rho(S_i, S) = \min_{S_i \in K_j, S_d \in CK_j} \rho(S_i, S_d) \right\}.$$

Объект $S \in B \cap K_j$, $j = 1, \dots, l$, принадлежит множеству шумовых объектов D_j класса K_j , если

$$\left\{ S_i \in E_0 \mid \rho(S_i, S) = \min_{S_i \in CK_j, S_d \in K_j} \rho(S_i, S_d) \right\} > \left\{ S_i \in K_j \mid \rho(S_i, S) < \min_{S_i \in K_j, S_d \in CK_j} \rho(S_i, S_d) \right\}. \quad (4)$$

- Для проверки условия (4) нужно определить:
- число объектов из CK_j , для которых $S \in B \cap K_j$ является ближайшим по метрике $\rho(x, y)$;
 - число выполненных неравенств вида $\rho(S_i, S) < \min_{S_i \in K_j, S_d \in CK_j} \rho(S_i, S_d)$ для объектов $S_i \in K_j$.

Поиск шумовых объектов по условию (4) имеет смысл, если число граничных объектов $|B|$ больше числа классов l . 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 $K_i \cap \left(E_0 \setminus \bigcup_{j=1}^l D_j\right)$, $i = 1, \dots, l$, разделены на минимальное число μ непересекающихся групп объектов $G_{i1}, \dots, G_{i\mu}$ по алгоритму из п. 1, $m_{ij} = |G_{ij}|$, $j = 1, \dots, \mu$, $\sum_{j=1}^{\mu} m_{ij} = m_i$. Для анализа результатов разбиения класса K_i на непересекающиеся группы с учетом их числа, представительности (по числу объектов) и удаления шумовых объектов предлагается использовать такую структурную характеристику, как оценка компактности

$$\Theta_i = \frac{\sum_{j=1}^{\mu} m_{ij}^2}{m_i^2}. \quad (5)$$

Очевидно, что множество допустимых значений Θ_i по выражению (5) лежит 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 проводится с учетом доли $\left(\frac{|E_0 \setminus \bigcup_{i=1}^l D_i|}{m}\right)$ исключенных из рассмотрения по условию (4) шумовых объектов как

$$R\left(E_0 \setminus \bigcup_{i=1}^l D_i, \rho\right) = \left(\frac{|E_0 \setminus \bigcup_{i=1}^l D_i|}{m}\right) \frac{\sum_{i=1}^l m_i \Theta_i}{|E_0 \setminus \bigcup_{i=1}^l D_i|} = \frac{\sum_{i=1}^l m_i \Theta_i}{m}. \quad (6)$$

Значения (5) и (6) косвенно свидетельствуют об однородности (неоднородности) струк-

туры обучающей выборки. Чем ближе сходство групп по числу входящих в них объектов класса, тем ближе значение (5) к $\frac{1}{m_i}$, а (6) — к $\frac{l}{m}$.

При $R\left(E_0 \setminus \bigcup_{i=1}^l D_i, \rho\right) = 1$ число групп объектов на $E_0 \setminus \bigcup_{j=1}^l D_j$ равно числу классов. Множество значений по (5) и (6) соответственно в $\left[\frac{1}{m_i}, 1\right]$ и $\left[\frac{l}{m}, 1\right]$ предлагается использовать в качестве меры компактности классов и выборки в целом.

Свойство связанности объектов классов по системе гипершаров, в пересечении которых содержатся объекты оболочек, позволяет получить разбиение выборки на минимальное число непересекающихся групп по алгоритму из п. 1. Такое разбиение в работе [5] предложено использовать для поиска минимального покрытия выборки объектами-эталоны по последовательности локальных областей признакового пространства, представленных объектами групп.

Обозначим $R_S = \rho(\bar{S}, S)$ расстояние от объекта $S \in K_t$ до ближайшего объекта $\bar{S}$ из противоположного к K_t класса ($\bar{S} \in CK_t$); δ — минимальное число непересекающихся групп из связанных и несвязанных объектов классов на E_0 .

Последовательность реализации процедуры поиска минимального покрытия объектами-эталоны обучающей выборки в работе [5] следующая. Упорядочим объекты каждой группы $G_u \cap K_t$, $u = 1, \dots, \delta$, $t = 1, \dots, l$, по множеству значений $\{R_S\}_{S \in G_u}$. В качестве меры расстояния между $S \in G_u$, $u = 1, \dots, \delta$, и произвольным допустимым объектом S' используется локальная метрика $d(S, S') = \rho(S, S')/R_S$. Решение о принадлежности S' к одному из классов $K_1, \dots, K_l$ или отказе от распознавания принимается по правилу: $S' \in K_t$, если

$$d(S_\mu, S') = \min_{S_j \in E_0} d(S_j, S') \text{ и } S_\mu \in K_t \\ \text{и } d(S_m, S') \neq \min_{S_j \in CK_t} d(S_j, S'). \quad (7)$$

Согласно принципу *последовательное исключение*, используемому 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 $E_{ed} = E_0$, $E_k = \emptyset$. Упорядочение по значениям отступа $\{R_S\}_{S \in G_u}$, $u = 1, \dots, \delta$, 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 правилу (7) остается корректным (без ошибок распознающим объекты) на E_0 .

Будем считать, что нумерация групп объектов отражает порядок $|G_1| \geq \dots \geq |G_\delta|$, и по группе G_p , $p = 1, \dots, \delta$, не проводился отбор объектов-эталонов. Кандидаты на удаление из E_{ed} последовательно выбираются начиная с $S \in G_p$, с минимальным значением R_S . Если включение $S \in E_k$ нарушает корректность решающего правила (7), то S возвращается в множество E_{ed} .

3. Отбор информативных наборов признаков

В работе [1] утверждалось, что при отборе информативных признаков с учетом компактности происходит невозрастание числа объектов-эталонов (столпов), необходимых для корректного распознавания объектов обучающей выборки. Для доказательства утверждения требовалось показать увеличение значения локальной компактности при удалении малоинформативных признаков, выражаемой через число объектов, притягиваемых столпами.

Очевидно, что число и состав шумовых объектов, определяемых условием (4), зависит от наборов признаков в описании объектов. Существует проблема согласования процессов отбора информативных признаков и удаления шумовых объектов. Для решения этой проблемы предлагается использовать алгоритм формирования информативных наборов признаков, состав которых зависит от выбора начальных приближений.

Пусть структура объектов классов на выборке E_0 вычисляется по алгоритму группировки из п. 1, минимальное покрытие объектами-эталонами с учетом удаления шумовых объектов получено по принципу *последовательное исключение* из п. 2. Обозначим $Sh(X(k))$ — число шумовых объектов E_0 , определяемых по условию (4) на наборе признаков $X(k) \subset X(n)$, CF — число объектов-эталонов минимального покрытия обучающей выборки, из которой удалены $Sh(X(k))$ шумовых объектов.

Идея поиска информативного набора $X(k) \subset X(n)$ заключается в отказе от перебора тех сочетаний признаков из $X(n)$, которые не ведут к оптимальному или локально-оптимальному решению. Обозначим

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

— мультипликативный показатель притягательной способности CF объектов-эталонов минимального покрытия с учетом числа $Sh(X(k))$ удаляемых из выборки шумовых объектов. Очевидно, что экстремальное значение

(8) равно $\frac{m}{l}$, которое достигается в случае отсутствия шумовых объектов и при одном эталоне на каждом из l классов.

Рассмотрим задачу поиска информативного набора $X(k) \subset X(n)$, $1 \leq k \leq n$, не прибегая к явному определению числа объектов-эталонов минимального покрытия CF . Обозначим P — подмножество индексов признаков из $X(n)$, $D_j(P)$ — множество шумовых объектов класса K_j , $j = 1, \dots, l$, по (4) на наборе $\{x_a\}_{a \in P}$.

Считается, что состав множества объектов в гипершаре

$$O(S_i, P) = \left\{ S \in K_j \mid \begin{aligned} &\rho(S, \bar{S}_i) < \rho(S_i, \bar{S}_i), \rho(S_i, \bar{S}_i) = \\ &\min_{S \in CK_i \cap \left(E_0 \setminus \bigcup_{t=1}^l D_t(P) \right)} \rho(S_i, S) \end{aligned} \right\}$$

с центром в $S_i \in E_0 \cap K_j$, $j = 1, \dots, l$, $\bar{S}_i \notin K_j$, и радиусом $\rho(S_i, \bar{S}_i)$ определяется по набору признаков $\{x_a\}_{a \in P}$. Оценку объекта $S_i \in E_0$ на $\{x_a\}_{a \in P}$ вычислим как

$$Z(S_i, P) = \begin{cases} 0, S_i \in \bigcup_{j=1}^l D_j(P); \\ |O(S_i, P)|. \end{cases} \quad (9)$$

Признак $x_d \in X(n)$ является кандидатом на включение в набор $\{x_a\}_{a \in P}$, если

$$\sum_{S_i \in E_0} Z(S_i, P \cup \{d\}) > \sum_{S_i \in E_0} Z(S_i, P).$$

Для отбора информативных наборов признаков предлагается алгоритм, описание пошаговой реализации которого приводится ниже.

Шаг 1. Выбор $i1, j1 \in \{1, \dots, n\}$, $i1 \neq j1$. $P = \{i1, j1\}$. $MAXP = 0$;

Шаг 2. Выделить $\bigcup_{j=1}^l D_j(P)$ по (4) на $\{x_a\}_{a \in P}$.
Вычислить $\{Z(S_i, P)\}_1^m$ по (9) на $\{x_a\}_{a \in P}$.

$$MAXP = \sum_{i=1}^m Z(S_i, P);$$

Шаг 3. $u = 0$. $C \stackrel{i=1}{=} 0$.

Для всех $v \in \{1, \dots, n\} \setminus P$

выделить $\bigcup_{j=1}^l D_j(P \cup \{v\})$ по (4)
на $\{x_a\}_{a \in P \cup \{v\}}$.

Оценки компактности минимального покрытия объектов по заданной метрике

Выборка данных	Базовая метрика	Число объектов		Значение (8)
		шумовых	эталонов	
Liver	Журавлева	183	51	5,0817
German	Журавлева	260	110	4,9782
Australian	Евклида	117	32	14,8700
Ionosphere	Евклида	23	16	19,1567

Единственность множества граничных объектов классов следует из структуры расстояний между объектами обучающей выборки, определяемой по заданной базовой метрике. Следствием из этого является единственность решения выбора числа и состава шумовых объектов по условию (4) и отбора эталонных объектов по принципу *последовательное исключение* из п. 2. Интегрирующим показателем между числом шумовых и эталонных объектов в табл. 1 является среднее число объектов (8), притягиваемых одним эталоном минимального покрытия выборки.

Решение по отбору информативных признаков алгоритмом из п. 3 не является устойчивым, так как оно зависит от выбора начального приближения. Какому из нескольких наборов отдать предпочтение, может быть определено по показателям обобщающей способности распознающих алгоритмов.

Обобщающая способность распознающих алгоритмов по правилу (7) на данных из работы [9] показана в табл. 2 на исходных и на некоторых из наборов информативных признаков. При вычислении показателей обобщающей

Таблица 2

Обобщающая способность алгоритмов с учетом удаления шумовых объектов

Характеристики	Выборка данных			
	Australian	Ionosphere	Liver	German
Информативный набор признаков	$x_2, x_3, x_5, x_8, x_{10}, x_{13}, x_{14}$	x_1, x_5, x_7, x_8	x_2, x_3, x_4, x_5, x_7	$x_1, x_2, x_3, x_4, x_5, x_8, x_{14}$
Точность на исходном наборе признаков	81,81	80,33	70,51	72,25
Точность на информативном наборе признаков	84,49	85,52	69,32	73,55

Вычислить $\{Z(S_i, P \cup \{v\})\}_1^m$ по (9)

на $\{x_a\}_{a \in P \cup \{v\}}$, $N = \sum_{i=1}^m Z(S_i, P \cup \{v\})$.

Если $N > C$, то $C = N$, $u = v$.

Шаг 4. Если $C > MAXP$, то $MAXP = C$, $P = P \cup \{u\}$, переход к шагу 2.

Шаг 5. Вывод P .

Шаг 6. Конец.

Состав наборов признаков и их число, определяемое алгоритмом, зависят от выбора значений множества P на первом шаге (индексов признаков $i1, j1$). Для выбора единственного решения из нескольких наборов предлагается использовать показатели обобщающей способности распознающих алгоритмов.

4. Вычислительный эксперимент

Методика отбора информативных признаков с учетом удаления шумовых объектов по условию (4) демонстрируется на четырех выборках данных из работы [11]. При вычислении множества граничных объектов классов были использованы метрика Евклида и метрика Журавлева

$$\rho(x, y) = \sum_{i \in I} |x_i - y_i| + \sum_{i \in J} \begin{cases} 1, & x_i \neq y_i; \\ 0, & x_i = y_i, \end{cases}$$

где I, J — множества индексов соответственно количественных и номинальных признаков. Для унификации масштабов измерений значения количественных признаков были отображены в $[0; 1]$.

Метрики Евклида и Журавлева рассматривали как базовые для формирования локальных метрик каждого объекта выборки. При определении локальной метрики объекта использовали расстояние до ближайшего (граничного) объекта из противоположного класса по базовой метрике. Шумовые объекты в условии (4) рассматриваются как подмножество граничных объектов классов. Удаление шумовых объектов приводит к изменению состава граничных объектов. Следствием из этого является изменение расстояний между объектами, вычисляемыми по локальным метрикам. В табл. 1 представлены оценки (8) компактности минимального покрытия выборок данных из работы [9] объектами-эталонами по заданной метрике при удалении шумовых объектов по условию (4).

Различия между алгоритмами

FRiS-Stolp	Новый алгоритм
Наличие шумовых объектов и их число определяются в процессе вычисления объектов-эталонов (столпов) выборки	Шумовые объекты выбираются из множества граничных объектов классов. Единственность выбора определяется по условию (4)
Выбор объекта $S \in E_0$ в качестве эталона проводится по критерию эффективности $R(S) = \lambda D_s + (1 - \lambda) T_s$, где D_s , T_s — характеристики соответственно оборотности и толерантности, параметр $\lambda \in [0,1]$	Выбор объектов-эталонов проводится на основе результатов разбиения объектов классов на непересекающиеся группы. Число и состав множества эталонов для корректного распознавания выборки определяются по каждой группе в отдельности

способности применяли 60 различных вариантов разделения объектов выборок на обучение и контроль в соотношении 9:1. Для удаления шумовых объектов использовали условие (4).

Набор информативных признаков, полученный алгоритмом из п. 3, не всегда является гарантией повышения обобщающей способности алгоритма распознавания на нем относительно исходного набора. Это видно на примере данных Liver из табл. 2. На исходном наборе обобщающая способность была 70,51, на информативном — 69,32. Изменение показателей компактности (8) на информативных наборах признаков из табл. 2 демонстрирует табл. 3.

Из сравнения результатов вычислительного эксперимента в табл. 1 и табл. 3 можно сделать следующие выводы. Существует различие между числом шумовых объектов и объектов-эталонов на исходном и информативном наборах признаков. Устойчивые изменения прослеживаются по показателю компактности (8). На информативных наборах признаков значения компактности выше, чем на исходных. Такое отношение ("выше") объясняется чувствительностью метрических алгоритмов к размерности пространства. Существуют наборы, добавление новых признаков в которые приводит к размыванию отношений сходства между объектами. Индикатором изменения отношений сходства служат значения показателей (8).

Использование оценок компактности позволяет ранжировать обучающие выборки в следующем порядке:

- 1) значение (8), равное $\frac{m}{l}$, определяет выборку, в которой нет шумовых объектов по условию (4) и каждый класс представлен одним объектом-эталонным минимального покрытия;
- 2) множество шумовых объектов, формируемое по условию (4), пусто, а компактность выборки по выражению (6) равна 1. Число объектов-эталонов минимального покрытия выборки больше числа классов;

Таблица 3

Оценки компактности минимального покрытия объектов на информативных наборах признаков из табл. 2

Выборка данных	Базовая метрика	Число объектов		Значение (8)
		шумовых	эталонов	
Liver	Журавлева	170	39	7,1041
German	Журавлева	284	88	5,8256
Australian	Евклида	120	24	19,6196
Ionosphere	Евклида	32	9	32,2130

3) множество шумовых объектов выборки, формируемое по условию (4), не пусто.

Изменение ранга обучающей выборки с 3 на 2 и с 2 на 1 может быть получено путем отбора информативных наборов признаков алгоритмом из п. 3.

Использованный для вычислительного эксперимента алгоритм распознавания является новым, и есть причины для сравнения его с известным алгоритмом построения решающего правила FRiS-Stolp [1,3]. Вычисление локальных метрик в алгоритмах проводится по одному принципу. Различия между алгоритмами представлены в табл. 4.

Как видно из табл. 4, для реализации нового алгоритма не требуется задавать дополнительные параметры. Процедуры отбора шумовых объектов и минимального покрытия выборки объектами-эталонами не пересекаются друг с другом. Сравнительный анализ двух алгоритмов (табл. 4) по показателю обобщающей способности является темой дальнейших исследований.

Заключение

Предложен способ выделения шумовых объектов как подмножества граничных объектов классов обучающей выборки по заданной базовой метрике. Описаны алгоритмы поиска минимального покрытия выборки объектами-эталонами и отбора информативных наборов признаков с учетом удаления шумовых объектов. Для селекции предложено использовать показатели компактности обучающих выборок и среднее число объектов, притягиваемых одним эталоном минимального покрытия выборки.

Список литературы

1. 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.
2. Субботин С. А. Комплекс характеристик и критериев сравнения обучающих выборок для решения задач диагностики и распознавания образов // Математичні машини і системи. 2010. № 1. С. 25–39.
3. Загоруйко Н. Г., Кутненко О. А. Цензурирование обучающей выборки // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 1(22). С. 66–73.
4. Борисова И. А., Кутненко О. А. Цензурирование ошибочно классифицированных объектов выборки // 17-я Всероссийская конференция "Математические методы распознавания образов — 2015". 19–25 сентября 2015, Россия, г. Светлогорск.
5. Игнатьев Н. А. Кластерный анализ данных и выбор объектов-эталонов в задачах распознавания с учителем // Вычислительные технологии. 2015. Т. 20, № 6. С. 34–43.
6. Згуральская Е. Н. Выбор информативных признаков для решения задач классификации с помощью искусственных нейронных сетей // Нейрокомпьютеры: разработка, применение. 2012. № 2. С. 20–27.
7. Saidov D. Y. Data visualization and its proof by compactness criterion of objects of classes // International Journal of Intelligent Systems and Applications (IJISA), 2017. Hang Kong, Vol 9, N. 8. P. 51–58.
8. Jeffery I., Higgins D., Culhane A. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data // BMC Bioinformatics, 2006. Vol. 7, P. 359.
9. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. Пер. с англ. Москва: ДМК Пресс, 2018. 652 с.
10. 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.
11. <http://archive.ics.uci.edu/ml/datasets>.

N. A. Ignatiev, D. Sc., Professor, e-mail: n_ignatev@rambler.ru,
National University of Uzbekistan named after M. Ulugbek, Tashkent

Compactness of Objects of Classes and Selection of Learning Samples

The relations between the sample objects $E_0 = \{S_1, \dots, S_m\}$, divided into disjoint classes $K_1, \dots, K_l$, is considered. Compactness determines the quantitative measure of these relations by a given metric $\rho(x, y)$ and various subsets of the set of features $X(n)$. The selection of the sample is based on the results of the sequential execution of three operators:

- search and delete noise objects;
- minimum sample coverage with objects—standards;
- selection informative set of features.

The rule according to which noise objects are considered as a subset of boundary objects of classes is described. A grouping algorithm based on the relationship " $\leftrightarrow$ " of connectedness objects of classes has been developed. For any pair (S_i, S_j) , $S_i, S_j \in K_d$, $d = 1, \dots, l$, included in one group, there always exists a chain $S_i \leftrightarrow S_k \leftrightarrow \dots \leftrightarrow S_j$. The search for the minimum sample coverage with reference objects is performed for each group separately. It is shown that when selecting informative features, the average number of objects attracted by one of the standard increases.

Keywords: noise object, informative features, selection of training samples, object—standard, generalizing ability, cross validation

DOI: 10.17587/it.25.545-552

References

1. Zagoruiko N. G., Kutnenko O. A., Zyryanov A. O., Levannov D. A. *Mashinnoe Obychenie i Analiz Dannih*, 2014, vol. 1, no. 7, pp. 891–901 (in Russian).
2. Subbotin S. A. *Matematichni Mashini i Sistemi*, 2010, no. 1, pp. 25–39 (in Russian).
3. Zagoruiko N. G., Kutnenko O. A. *Vestnik Tomskogo Gosudarstvennogo Universiteta. Upravleniye, Vychislitel'naya Tekhnika*, 2013, no. 1(22), pp. 66–73 (in Russian).
4. Borisova I. A., Kutnenko O. A. *Tsenzurirovaniye oshibochno klassifitsirovannykh ob'yektov vyborki* (The erroneous classify objects in dataset censoring), *17-ya Vserossiyskaya konferentsiya "Matematicheskiye metody raspoznavaniya obrazov — 2015"*, Svetlogorsk, 2015 (in Russian).
5. Ignat'ev N. A. *Vychislitel'nyye Tekhnologii*, 2015, no. 6, pp. 34–43 (in Russian).
6. Zguralskaya E. N. *Neyrokomp'yutery: Razrabotka, Prime-neniye*, 2012, no. 2, pp. 20–27 (in Russian).
7. Saidov D. Y. *International Journal of Intelligent Systems and Applications (IJISA)*, 2017, vol. 9, no. 8, pp. 51–58.
8. Jeffery I., Higgins D., Culhane A. *BMC Bioinformatics*, 2006, vol. 7, pp. 359.
9. Goodfellow I., Bengio Y., Courville A. *Deep Learning*, Moscow, DMK Press, 2018, 652 p.
10. Aivazyan S. A., Buchstaber V. M., Yenyukov I. S., Meshalkin L. D. *Applied statistics. Classification and reduction of dimensionality*, Moscow, Finansy i statistika, 1989, 608 p. (in Russian).
11. Available at: <http://archive.ics.uci.edu/ml/datasets>