УДК 519.95 DOI: 10.17587/it.27.18-24

Н. А. Игнатьев, д-р физ.-мат. наук, проф., e-mail: ignatev@rambler.ru,
М. Я. Лолаев, преподаватель, e-mail: musulmon.lolayev.94@mail.ru,
Национальный университет Узбекистана, г. Ташкент

Анализ соответствия структур отношений объектов классов на многообразиях их описаний

Рассматривается синтез латентных признаков в целях снижения размерности пространства для описания объектов непересекающихся классов. Латентные признаки синтезируются на данных, полученных как результат предобработки значений исходных признаков. Исследуется соответствие структур отношений объектов на плоскости по результатам предобработки и группировки данных. Предложена мера соответствия двух структур отношений объектов в разных признаковых пространствах.

***Ключевые слова**: предобработка данных, метод главных компонент, мера соответствия, снижение размерности*

Введение

Рассматривается анализ соответствия структур отношений на многообразиях их описаний в задачах снижения размерности признакового пространства. В работе [1] выделены следующие типовые прикладные задачи снижения размерности:

— отбор наиболее информативных признаков (включая латентные);

— сжатие массивов обрабатываемой и хранимой информации;

— визуализация (наглядное представление) данных;

— построение условных координатных осей (многомерное шкалирование, латентно-структурный анализ).

Существуют несколько постановок задачи выбора пространства из латентных признаков для описания объектов. Одним из существенных различий между ними является наличие или отсутствие классификации объектов. Метод главных компонент (PCA) принадлежит к числу наиболее часто используемых при анализе данных [1]. Например, на основе PCA разработан метод локальной геометрии [2] для поиска скрытых закономерностей в данных. Линейное отображение описаний объектов на плоскость позволило экспертам визуально исследовать структуру отношений между ними.

Искажение структуры данных при проецировании в пространство меньшей размерности связано с изменением отношений близости между объектами [3]. Были предложены два способа для сравнения структур данных до и после проецирования их на двумерную плоскость.

Для сравнения первым способом в каждой точке на плоскости вычисляется множество K1 ближайших к данной точке соседних точек в исходном пространстве и множество K2 ближайших соседей в двумерном пространстве после проецирования. Мерой сохранения отношений соседства между точками данных после их проецирования служит мощность пересечения K1 и K2 (число совпадающих точек).

Второй способ для анализа соответствия структур отношений объектов в разных признаковых пространствах использует нейронные сети. Отношение соседства определяется по узлам сетки в самоорганизующих картах Кохонена в исходном признаковом пространстве и пространстве меньшей размерности. Узлы сетки образовывают прямоугольную или гексагональную структуру.

К недостаткам двух приведенных выше способов для анализа структур отношений из работы [3] можно отнести следующие:

— при выборе меры расстояния для сравнения объектов не учитываются масштабы измерений признаков;

— не определены правила выбора числа ближайших соседей K1 и K2 для вычисления меры сохранения отношений соседства;

— нет четких, обоснованных рекомендаций по выбору узлов сетки для самоорганизующих карт Кохонена.

Существенное значение для формирования структуры отношений между объектами имеет выбор способа нормирования данных. Одной из целей нормирования является инвариантность к масштабам измерений признаков. Свойство инвариантности расширяет возможности для обнаружения скрытых закономерностей, характерных для всех выборок данных из генеральной совокупности.

Важным показателем для анализа данных является оценка компактности объектов классов. Единой меры компактности не существует [4]. Различаются вычисления мер компактности в зависимости от размерности данных. Экстремум критерия для разбиения значений признака на непересекающиеся интервалы из работы [5] может рассматриваться как мера компактности, инвариантная к масштабам измерения. Перечень факторов, от которых зависит структура отношений объектов при размерности пространства, большей либо равной 2, приводится в работе [6].

Применение методов анализа данных без каких-либо предположений о природе их среды рассматривается как средство обнаружения скрытых закономерностей в слабо структурированных предметных областях. Исследование отличия истинных параметров среды от полученных на основе предположений проводится в целях повышения адекватности описания реальных процессов и явлений в рамках математических моделей. В работе [7] показано, что порог линейной решающей функции, вычисленный по экстремуму критерия разбиения признаков на непересекающиеся интервалы [5], отличается от порога, вычисляемого при предположении о нормальном распределении данных в линейном дискриминанте Фишера.

Синтез латентных признаков может проводиться как по одному набору признаков, так и по разным. Например, в работе [8] каждый латентный признак формируется по правилам иерархической агломеративной группировки по "*своему*" набору исходных признаков. Результаты использования метода главных компонент являются примером синтеза латентных признаков из одного набора.

Для решения ряда задач выбора латентного признакового пространства существуют доступные для использования инструментальные средства. К таковым можно отнести функции из библиотеки языка PYTHON [9]. Для оценки структуры отношений объектов в латентном признаковом пространстве рекомендуется использовать меру компактности объектов классов из работы [6].

Значения латентных признаков зависят от предобработки данных в исходном признаковом пространстве. Эта зависимость отражается на структуре отношений объектов в латентном признаковом пространстве. Требуется разработка мер соответствия структур отношений, при формировании которых применялись разные способы предобработки. Меры соответствия структур могут быть востребованы при безпризнаковом распознавании, например, при использовании метода Саймона [10] для визуализации объектов в R^2 по матрице их парных расстояний.

В данной работе визуализация проводится в целях:

— анализа соответствия структур отношений объектов в проекциях на плоскость при разных способах предобработки исходных данных;

— выбора представления данных в пространстве меньшей размерности со значением меры компактности объектов классов больше, чем в исходном.

Поиск подходящей размерности пространства для процедуры проецирования в пространство более низкой размерности на основе меры компактности объектов классов предлагается проводить по упорядоченной последовательности наборов исходных (сырых) признаков. Последовательность наборов [11] формируется по значениям отношений внутриклассового сходства и межклассового различия признаков.

1. Постановка задачи

Рассматривается множество $E_0 = \{S_1, ..., S_m\}$ объектов, разделенное на два непересекающихся подмножества (класса) K_1 и K_2. Описание объектов проводится с помощью набора из n количественных признаков $X(n) = (x_1, ..., x_n)$. На $X(h) \subset X(n)$, $2 \leqslant h \leqslant n$, определены операторы $A_1, ..., A_\mu$, $\mu \geqslant 2$, для предобработки данных с сохранением размерности признакового пространства. Для снижения размерности признакового пространства описания объектов E_0 из $X(h)$ отображаются в R^2. Исследуется зависимость топологии объектов в R^2 от применения операторов $\{A_i\}$, $i = 1, ..., \mu$.

Компактность объектов классов определяется через структуру их отношений в призна-

ковом пространстве по заданной мере расстояния. Для оценки компактности предлагается использовать отношение связанности объектов по множеству граничных объектов классов.

Считается, что задана метрика $\rho(x, y)$ для анализа отношений по описанию объектов E_0. Обозначим

$$L(E_0,\rho) = \{S \in E_0 \,|\, S_i \in K_{3-t}, \rho(S,S_i) = \\ = \min_{S_j \in K_t} \rho(S_j,S_i), t = 1,2\}$$

— множество граничных объектов классов, определяемое на E_0 по метрике $\rho(x, y)$. Объекты S_i, $S_j \in K_t$, $t = 1, 2$, считаются связанными между собой ($S_i \leftrightarrow S_j$), если $\{S \in L(E_0, \rho) | \rho(S, S_i) < r_i$ и $\rho(S, S_j) < r_j\} \neq \varnothing$, где $r_i(r_j)$ — расстояние до ближайшего от $S_i(S_j)$ объекта из K_{3-t} по метрике $\rho(x, y)$.

Множество $G_{tv} = \{S_{v_1}, \ldots, S_{v_c}\}$, $c \geqslant 2$, $G_{tv} \subset K_t$, $v < |K_t|$ представляет область (группу) со связанными объектами в классе K_t, если для любых $S_{v_i}, S_{v_j} \in G_{tv}$ существует путь $S_{v_i} \leftrightarrow S_{v_k} \leftrightarrow \ldots \leftrightarrow S_{v_j}$. Объект $S_i \in K_t$, $t = 1, 2$, 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 $S_j \in K_t$.

Алгоритм определения минимального числа непересекающихся групп из связанных и несвязанных объектов по каждому классу K_t, $t = 1, 2$, описан в работе [6]. Доказана единственность разбиения по числу групп и составу объектов в них. Требуется:

— отобразить описание объектов E_0 из $X(h)$ в R^2 с использованием предобработки данных операторами $\{A_i\}$, $i = 1, \ldots, \mu$;

— оценить соответствие структур отношений объектов из E_0 в R^2 после предобработки данных операторами A_i, A_j, $i \neq j$.

2. О предобработке данных на основе классификации

Под предобработкой данных далее будем понимать преобразование значений признаков из $X(n)$ с учетом разделения объектов на непересекающиеся классы. Предобработка данных предшествует реализации алгоритмов снижения размерности пространства через синтез латентных признаков для описания объектов. Интерес для исследования представляют:

— преобразования признаков, инвариантные к масштабам их измерений;

— выбор условия для последовательного формирования наборов признаков $X(h) \subset X(n)$, $2 \leqslant h \leqslant n$;

— поиск максимального значения меры компактности объектов классов в R^2 при отображении их описаний из $X(h)$ в R^2.

Инвариантность к масштабам измерений определяется через оценку компактности значений признака в двухклассовой задаче распознавания. Оценка компактности вычисляется как экстремум критерия при разбиении значений признака на непересекающиеся интервалы. В основе вычислений лежит проверка истинности гипотезы "*каждый интервал содержит значения признака всех объектов одного класса*".

Пусть u_i^1, u_i^2 — число значений признака x_j класса K_i, $i = 1,2$, соответственно в интервалах $[c_1, c_2]$, $(c_2, c_3]$, $|K_i| > 1$, t — порядковый номер элемента упорядоченной по возрастанию последовательности $r_{j_1}, \ldots, r_{j_t}, \ldots, r_{j_m}$ значений x_j у объектов из E_0, определяющий границы интервалов как $c_1 = r_{j_1}$, $c_2 = r_{j_t}$, $c_3 = r_{j_m}$. Критерий

$$\left(\frac{\sum_{i=1}^{2} u_i^1(u_i^1 - 1) + u_i^2(u_i^2 - 1)}{\sum_{i=1}^{2} |K_i|(|K_i| - 1)} \right) \times \\ \times \left(\frac{\sum_{d=1}^{2} \sum_{i=1}^{2} u_i^d \left(|K_{3-i}| - u_{3-i}^d \right)}{2|K_1||K_2|} \right) \rightarrow \max_{c_1 < c_2 < c_3} \quad (1)$$

позволяет вычислять оптимальное значение границы c_2 для интервалов $[c_1, c_2]$ и $(c_2, c_3]$.

Экстремум критерия (1) используется в качестве веса v_j ($0 < v_j \leqslant 1$) признака x_j. При $v_j = 1$ значения признака x_j у объектов из классов K_1 и K_2 не пересекаются между собой.

Реализация оператора для предобработки данных, инвариантных к масштабам измерений, разделяется на два этапа:

— вычисление границ двух непересекающихся интервалов для признака $x_i \in X(n)$ по критерию (1) на E_0. Выбор экстремального значения критерия (1) в качестве веса признака;

— нормирование признака $x_i \in X(n)$ по значениям границ интервалов и умножение на его вес.

Предобработка значений признака $x_j \in X(n)$ с учетом разбиения на интервалы $[c_1, c_2]$, $(c_2, c_3]$ по критерию (1) выглядит следующим образом:

$$x_j^* = v_j \frac{x_j - c_2}{c_3 - c_1}. \quad (2)$$

Для вычисления весов признаков, значения которых в отличие от критерия (1) зависят от принадлежности к набору $X(h) \subset X(n)$, $2 \leqslant h \leqslant n$, используется функционал [11]

$$J(w) = \frac{\sum_{i=1}^{h} w_i \theta_i}{\sum_{i=1}^{h} w_i \gamma_i} \to \min, \qquad (3)$$

где θ_i, γ_i — значения соответственно внутриклассовой близости и межклассового различия по признаку $x_i \in X(h)$. Функционал имеет бесконечное множество решений. При ограничении на веса $\sum_{i=1}^{h} w_i = 1, w_i \geqslant 0$ преобразуем (3) к виду

$$F(w, \lambda) = \frac{\sum_{i=1}^{h} w_i \theta_i}{\sum_{i=1}^{h} w_i \gamma_i} + \lambda \left(\sum_{i=1}^{h} w_i - 1 \right). \qquad (4)$$

Используя метод неопределенных множителей Лагранжа к соотношению (4), получим

$$w_i = \begin{cases} \dfrac{\gamma_i - \theta_i}{\sum\limits_{\{j | \gamma_j - \theta_j > 0\}} \gamma_j - \theta_j}, \gamma_i - \theta_i > 0, \\ 0, \gamma_i - \theta_i \leqslant 0. \end{cases} \qquad (5)$$

Множество допустимых значений весов $\{w_i\}$ по выражению (5) зависит от способов вычисления $\{\theta_i, \gamma_i\}$. Рассмотрим два таких способа.

1 способ. Пусть m_{1i} и m_{2i} — математические ожидания признака $x_i \in X(n)$ для объектов из классов K_1 и K_2. Тогда

$$\theta_i = \sum_{t=1}^{2} \sum_{S_j \in K_t} \left| x_{ji} - m_{ti} \right| / m,$$
$$\gamma_i = \sum_{t=1}^{2} \sum_{S_j \in K_{3-t}} \left| x_{ji} - m_{ti} \right| / m. \qquad (6)$$

2 способ. Преобразуем значения количественного признака $x_i \in X(n)$ в два значения (градации) в номинальной шкале измерений. Выбор градаций проводится по двум интервалам $[c_1, c_2]$ и $(c_2, c_3]$, границы которых определены по соотношению (1). Обозначим g_{1i}^t, g_{2i}^t — частоты встречаемости значений градации t ($t = 1, 2$) в описании объектов в K_1 и K_2. Тогда

$$\gamma_i = 1 - \frac{\sum_{t=1}^{2} g_{1i}^t g_{2i}^t}{|K_1||K_2|};$$
$$\theta_i = 1 - \frac{\sum_{t=1}^{2} g_{1i}^t (g_{1i}^t - 1) + g_{2i}^t (g_{2i}^t - 1)}{|K_1|(|K_1| - 1) + |K_2|(|K_2| - 1)}. \qquad (7)$$

Вычисление внутриклассового сходства и межклассового различия по соотношению (7) востребовано при описании объектов в разнотипном признаковом пространстве. При вычислении значений g_{1i}^t, g_{2i}^t нужно учитывать, что числа градаций у номинальных признаков разные и не обязательно равны 2.

Множество значений $\{\theta_i, \gamma_i\}$ можно использовать при формировании набора признаков $X(h) \subset X(n)$, $2 \leqslant h < n$. Согласно выводам теоремы из работы [11] признак x_j является кандидатом на удаление из $X(h + 1)$, если

$$\frac{\theta_j}{\gamma_j} = \max_{x_i \in X(h+1)} \frac{\theta_i}{\gamma_i}. \qquad (8)$$

В силу условия $\sum_i w_i = 1$ значение весов признаков (5) на каждом наборе $X(h)$ будут разные.

3. О соответствии структур отношений объектов классов при предобработке данных

Пусть после предобработки данных по набору $X(h)$, $2 \leqslant h \leqslant n$, оператором A_i, $i = 1, ..., \mu$, проведено линейное отображение описаний объектов E_0 в R^2. При анализе структуры данных по отношению связанности объектов классов в R^2 получено разбиение объектов на множество непересекающихся групп $Z_i = \{G_{id}\}_{d=1}^{p(i)}$, $2 \leqslant p(i) \leqslant m$. Для сравнения двух структур отношений объектов в R^2 по Z_i и Z_j, $i \neq j$, сформируем матрицу $D = \{d_{uv}\}$, $u = 1,...,m$, $v = 1, 2, 3$. Значениями элементов d_{r1}, d_{r2} являются номера групп, к которым принадлежит объект $S_r \in E_0$ соответственно в Z_i и Z_j, d_{r3} — номер области, содержащей пересечение этих групп.

Обозначим $p_1(i)$, $p_2(i)$ — число непересекающихся групп объектов из классов K_1 и K_2 в Z_i, $p(i) = p_1(i) + p_2(i)$. Максимальное число пересечений по Z_i и Z_j ограничено сверху:

$$p_1(i)\, p_1(j) + + p_2(i)\, p_2(j).$$

Пусть η — число пересечений групп из Z_i и Z_j, содержащих q_1, ..., q_η объектов, $\sum_{k=1}^{\eta} q_k = m$. Мера сходства структур отношений объектов по Z_i и Z_j будет вычисляться как

$$\Omega(Z_i, Z_j) = \frac{2 \sum_{k=1}^{\eta} (q_k)^2}{\sum_{k=1}^{p(i)} |G_{ik}|^2 + \sum_{k=1}^{p(j)} |G_{jk}|^2}. \qquad (9)$$

Максимальное значение меры сходства (9), равное 1, получается при $\eta = p(i) = p(j)$.

Меру сходства (9) можно использовать для анализа соответствия структур отношений объектов в $X(h)$, $2 \leqslant h < n$, и в R^2 или в $X(h)$ и $X(t)$ при $t \neq h$.

Для оценки качества отображений описаний объектов из $X(h)$ в R^2 предлагается использовать меру компактности, полученную через отношение связанности объектов. В качестве меры расстояния в R^2 рассматривается евклидова метрика. Меру компактности объектов по Z_i в классе K_t, $t = 1, 2$, определим как

$$\Omega(Z_i, K_t) = \frac{\sum\limits_{\{G_{id} \in Z_i | G_{id} \cap K_t \neq \varnothing\}} |G_{id}|^2}{|K_t|^2}, \qquad (10)$$

а в целом по выборке E_0 как

$$\Omega(Z_i, K_1 \cup K_2) = \\ = \frac{|K_1|\Omega(Z_i, K_1) + |K_2|\Omega(Z_i, K_2)}{m}. \qquad (11)$$

Дополнительным показателем при анализе структуры отношений объектов классов является число шумовых объектов. В данной работе множество шумовых объектов рассматривается как подмножество граничных объектов классов $L(E_0, \rho)$ по евклидовой метрике. Объект $S \in L(E_0,\rho) \cap K_j$, $j = 1, 2$, принадлежит множеству шумовых объектов D_j класса K_j, если

$$\left|\left\{S_i \in E_0 \left| \rho(S_i, S) = \min_{S_i \in K_{3-j}, S_d \in K_j} \rho(S_i, S_d)\right.\right\}\right| > \\ > \left|\left\{S_i \in K_j \left| \rho(S_i, S) < \min_{S_i \in K_j, S_d \in K_{3-j}} \rho(S_i, S_d)\right.\right\}\right|. \qquad (12)$$

Удаление множеств D_1 и D_2 из E_0 изменяет структуру отношений объектов на $E_0 \backslash (D_1 \cup D_2)$ и служит средством повышения обобщающей способности алгоритмов распознавания по правилу ближайшего соседа.

4. Эксперименты

Вычислительный эксперимент проводили на выборке, содержащей данные о поражениях желудочно-кишечного тракта из работы [12]. Выборка представлена 76 объектами, разделенными на класс K_1 (доброкачественные поражения) и K_2 (злокачественные поражения), $|K_1| = 21$, $|K_2| = 55$. Каждый объект описывался 1394 количественными признаками.

При формировании последовательности из наборов признаков $X(n - 1)$, $X(n - 2)$, ..., $X(2)$ использовалось условие (8). Значение внутриклассового сходства θ_j и межклассового различия γ_j в условии (8) вычислялось по выражению (6). Для программной реализации метода РСА использовались функции из библиотеки языка PYTHON [13].

Пусть $R2(h)$ обозначает набор признаков, полученный при проецировании описаний объектов E_0 из $X(h)$ в R^2. Исследовалась зависимость показателей компактности объектов классов (10) и выборки в целом (11) по $R2(h)$ от результатов предобработки данных по набору $X(h)$. Рассматривались следующие варианты представления данных:

— без предобработки;

— с нормированием в [0; 1];

— с преобразованием по соотношению (2).

Результаты анализа данных по трем вариантам в $R2(h)$ и мера сходства структур отношений объектов в $X(h)$ и в $R2(h)$ приводится в табл. 1—3, визуальное представление объектов в R^2 представлено на рис. 1—3.

Таблица 1

Анализ сходства структур отношений объектов по $X(h)$ и $R2(h)$ без предобработки данных

Размерность h	Число шумовых объектов по соотношению (12)	Компактность по выражениям (10), (11)			Сходство по соотношению (9)
		K_1	K_2	$K_1 \cup K_2$	
1394	13	0,1565	0,4010	0,3334	0,6820
806	18	0,1882	0,4023	0,3432	0,6935
122	18	0,1882	0,4023	0,3432	0,6935
9	18	0,1791	0,4056	0,3430	0,7067

Таблица 2

Анализ сходства структур отношений объектов по $X(h)$ и $R2(h)$ с учетом нормирования данных в [0;1]

Размерность h	Число шумовых объектов по соотношению (12)	Компактность по выражениям (10), (11)			Сходство по соотношению (9)
		K_1	K_2	$K_1 \cup K_2$	
1394	19	0,1338	0,4519	0,3640	0,6309
806	24	0,1111	0,3408	0,2773	0,0159
122	17	0,3151	0,7051	0,5973	0,0197
9	12	0,6054	0,8611	0,7904	0,9831

Таблица 3

Анализ сходства структур отношений объектов по $X(h)$ и $R2(h)$ с предобработкой данных по (2)

Размерность h	Число шумовых объектов по соотношению (12)	Компактность по выражениям (10), (11)			Сходство по соотношению (9)
		K_1	K_2	$K_1 \cup K_2$	
1394	16	0,4059	0,6516	0,5837	0,8441
806	13	0,4649	0,6826	0,6225	0,8608
122	15	0,8231	0,8288	0,8272	0,1010
9	12	0,4693	0,8280	0,7289	0,9474

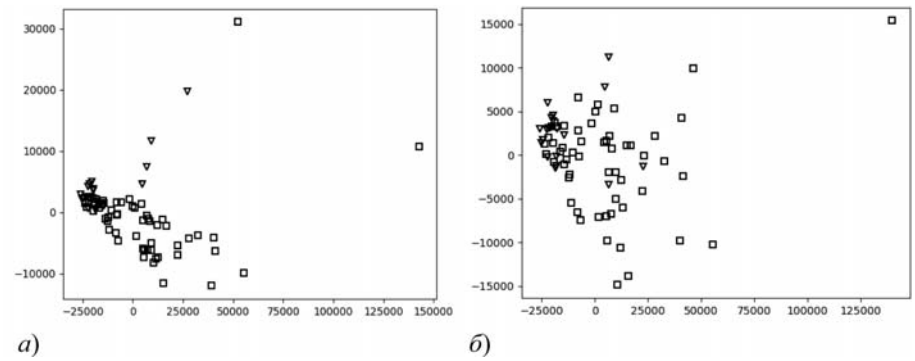

Рис. 1. Визуализация объектов без предобработки данных в пространстве:
а — R2(1394); *б* — R2(122)

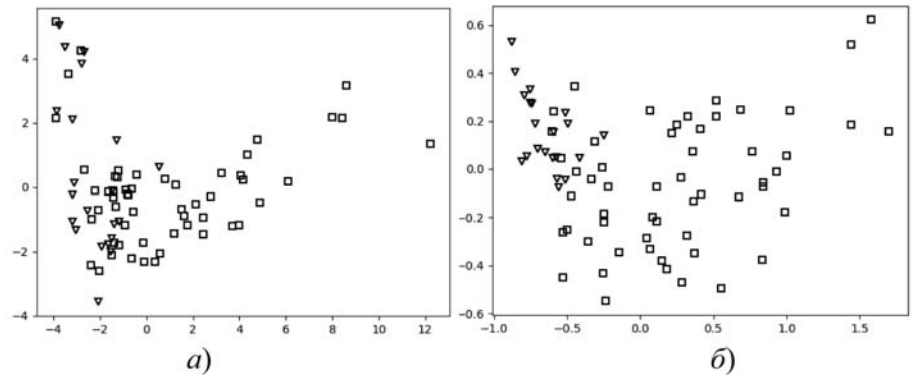

Рис. 2. Визуализация объектов с нормированием данных в [0;1] в пространстве:
а —R2(1394); *б* — R2(9)

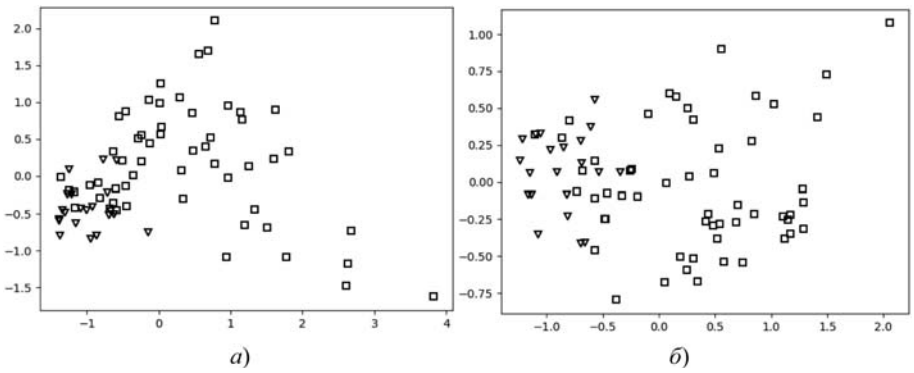

Рис. 3. Визуализация объектов при предобработке данных по соотношению (2) в пространстве:
а — R2(1394); *б* — R2(122)

Анализ табл. 1—3 показывает целесообразность использования предобработки данных. Показатели компактности объектов при нормировании данных в [0; 1] выше, чем на данных без предобработки. Самые высокие показатели компактности получены по $R2(122)$ при предобработке данных по соотношению (2).

Существует прямая зависимость между мерой компактности объектов классов и обобщающей способностью алгоритмов распознавания по правилу ближайшего соседа. Доказательство этого утверждения в форме вычислительного эксперимента приводится в работе [6].

Проверка соответствия структур отношений по $R2(h)$ без предобработки и с использованием предобработки данных представлен в табл. 4.

Анализ значений меры соответствия из табл. 4 показывает, что предобработка данных сильно искажает структуру отношений объектов. Объясняется это тем, что использование предобработки данных приводят к повышению компактности объектов классов в R^2.

Заключение

Рассмотрена проблема снижения размерности признакового пространства и ее связь с предобработкой данных. Способы предобработки рассчитаны на использование таких операций, как нормирование признаков, умножение значения признака на его вес. Разработана мера соответствия структур отношений объектов в разных признаковых пространствах. При вычислении меры использовалось разбиение выборки на непересекающиеся группы. Для группировки применялось отношение связанности объектов классов по определяемому множеству гипершаров. Обоснованием выбора способа предобработки данных служит значение меры компактности классов и выборки в целом.

Таблица 4

Соответствие структур отношений объектов в $R2(h)$ без предобработки и с предобработкой данных

Размерность h	Мера соответствия по (9) с учетом предобработки	
	нормированием в [0, 1]	преобразованием по (2)
1394	0,6309	0,8440
806	0,5783	0,0243
122	0,0112	0,0394
9	0,0199	0,0185

Список литературы

1. **Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.** Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.
2. **Дюк В. А.** Методология поиска логических закономерностей в предметной области с нечеткой системологией: На примере клин.-эксперим. исслед. Дисс. д-ра техн. наук. Санкт-Петербург, 2005. 309 с.
3. **Зиновьев А. Ю.** Визуализация многомерных данных. Красноярск: Изд. КГТУ, 2000.
4. **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.
5. **Zguralskaya E. N.** Analysis of the structure of the relationship between the descriptions of objects of classes and evaluation of their compactness // Workshop Proceedings Information Technology and Nanotechnology (ITNT-2019). Samara, Russia, May 21—24. 2019. P. 283—289. URL: http:/ceur-ws.org/Vol-2416.
6. **Ignatyev N. A.** Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis. 2018. Vol. 28, N.4. P. 590—597.
7. **Саидов Д. Ю.** Информационные модели на основе нелинейных преобразований признакового пространства в задачах распознавания. Дисс. доктора философии (PhD) по физико-математическим наукам. Ташкент, 2017. 93 с.
8. **Saidov D. Y.** Data visualization and its proof by compactness criterion of objects of classes // International Journal of Intelligent Systems and Applications (IJISA). 2017. Vol. 9, N. 8. P. 51—58.
9. **Мюллер А., Гвидо С.** Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. СПб.: ООО "Альфа-книга", 2017. 480 с.
10. **Мясников Е. В.** Анализ методов снижения размерности в задаче представления коллекции цифровых изображений // Компьютерная оптика. 2008. Т. 32, № 3. С. 296—301.
11. **Игнатьев Н. А.** Выбор минимальной конфигурации нейронных сетей // Вычислительные технологии. 2001. Т. 6, № 1. С. 23—28.
12. **Mesejo P.** et al. Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy // IEEE Transactions on Medical Imaging. 2016. Vol. 35, N. 9. P. 2051—2063.
13. **Scikit**-*le*arn user guide: Release 0.21.3. 2019.

N. A. Ignat'ev, D. Sc., Professor, e-mail: ignatev@rambler.ru,
M. Y. Lolaev, Teacher, musulmon.lolayev.94@mail.ru,
National University of Uzbekistan, Tashkent

An Analysis of the Compliance of the Structures of Relations of Objects of Classes on the Varieties of their Descriptions

The synthesis of latent features is considered in order to reduce the size of the space for describing objects of disjoint classes K_1, K_2. A condition is proposed for the sequential formation of feature sets $X(h) \subset X(n)$, $2 \leqslant h \leqslant n$ for linear display of object descriptions from $X(h)$ to R^2. Latent features are synthesized from data obtained as a result of preprocessing the values of the initial features. Preprocessing is realized through normalization in [0,1] and data transformation that is invariant to the scales of feature measurements. The correspondence structures of objects in $X(h)$ and R^2 are investigated according to the results of data preprocessing and grouping. A measure of compliance of two structures of relations of objects in different features spaces is proposed. The measure of compliance is calculated by splitting the objects of classes into disjoint groups. When splitting into groups, the relation of connectedness of class objects by a defined set of hyper balls is used.

***Keywords**: data preprocessing, principal component analysis, compactness measure, dimensionality reduction*

DOI: 10.17587/it.27.18-24

References

1. **Ayvazyan S. A., Bukhshtaber V. M., Yenyukov I. S., Meshalkin L. D.** Prikladnaya statistika. Klassifikatsiya i snizheniye razmernosti, Moscow, Finansy i statistika, 1989, 607 p. (in Russian).
2. **Dyuk V. A.** Metodologiya poiska logicheskikh zakonomernostey v pred-metnoy oblasti s nechetkoy sistemologiyey: na primere klin.-eksperim. Issled, Dis. d-ra tekhn. nauk. Sankt-Peterburg, 2005, 309 p. (in Russian).
3. **Zinov'yev A. Yu.** Vizualizatsiya mnogomernykh dannykh, Krasnoyarsk, Izd. KGTU, 2000 (in Russian).
4. **Zagoruyko N. G., Kutnenko O. A., Zyryanov A. O., Levanov D. A.** Obucheniye raspoznavaniyu obrazovannykh bez pereobucheniya, *Mashinnoye Obucheniye i Analiz Dannykh*, 2014, vol. 1, no. 7, pp. 891—901 (in Russian).
5. **Zguralskaya E. N.** Analysis of the structure of the relationship between the descriptions of objects of classes and evaluation of their compactness, *Workshop Proceedings Information Technology and Nanotechnology (ITNT-2019),* Samara, Russia, May 21—24, 2019, pp. 283—289, available at: http:/ceur-ws.org/Vol-2416 (in Russian).
6. **Ignatyev N. A.** Structure Choice for Relations between Objects in Metric Classification Algorithms, *Pattern Recognition and Image Analysis*, 2018, vol. 28, no. 4, pp. 590—597.
7. **Saidov D. Yu.** Informatsionnyye modeli na osnove nelineynogo preobladaniya priznakovogo prostranstva v zadachakh raspoznavaniya, Diss. doktor filosofii (PhD) po fiziko-matematicheskim naukam. Tashkent, 2017, 93 p. (in Russian).
8. **Saidov D. Y.** Data visualization and its proof by compactness criterion of objects of classes, *International Journal of Intelligent Systems and Applications (IJISA),* 2017, vol. 9, no. 8, pp. 51—58.
9. **Muller A., Gvido S.** Vvedeniye v mashinnoye obucheniye s pomoshch'yu Python. Rukovodstvo dlya spetsialistov po rabote s dannymi, SPb., OOO "Al'fa-kniga", 2017, 480 p. (in Russian).
10. **Myasnikov Ye. B.** Analiz metodov umen'sheniya razmera v zadache predstavleniya kollektsii tsifrovykh izobrazheniy, *Komp'yuternaya Optika*, 2008, vol. 32, no. 3, pp. 296—301 (in Russian).
11. **Ignat'yev N. A.** Vybor minimal'noy konfiguratsii neyronnykh setey, *Vychislitel'nyye Tekhnologii*, Novosibirsk, 2001, vol. 6, no. 1, pp. 23—28 (in Russian).
12. **Mesejo P.** et al. Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy, *IEEE Transactions on Medical Imaging*, 2016, vol. 35, no. 9, pp. 2051—2063.
13. **Scikit**-learn user guide: Release 0.21.3, 2019.