СВЯЗНОСТЬ ОБЪЕКТОВ КЛАССОВ И ЕЕ ВЛИЯНИЕ НА ОБОБЩАЮЩУЮ СПОСОБНОСТЬ АЛГОРИТМОВ

Турсунмуротов Д.Х.
Старший преподаватель кафедры технологий программирования факультета программной инженерии Ташкентского университета информационных технологий имени Мухаммада аль-Хорезми

Annotatsiya: Рассматриваются методы анализа структуры отношений классифицированных объектов в пространстве разнотипных признаков. Оценка отношений проводилось по мере компактности и через визуальное представления объектов на плоскости.

Ключевые слова: Компактность, визуализация, DBSCAN.

Для оценки структуры обучающей выборки разработаны различные методы, принципы их работы существенно отличаются друг от друга. 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 $E_0=\{S_1,\dots,S_m\}$ объектов, разделённое на $l(l>2)$ непересекающихся подмножеств (классов) $K_1, \dots, K_l$, $E_0=\bigcup_{i=1}^{l} K_i$. Описание объектов производится 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. 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 $S_i,S_j \in K_t$, $t=1,\dots,l$ 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 CK_t ($CK_t=E_0 \backslash K_t$) по метрике $\rho(x,y)$.

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

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

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

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 $\rho(S_i, S_{i_\tau}) < r_i$, $\tau = 1,\dots,\beta\text{-}1$. В $O(S_i)$ всегда существует непустое подмножество объектов

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

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

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

(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 $G_1,\dots,G_\theta$ таково.

Шаг 1: $\theta=0$;

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

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

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

Шаг 5: Конец.

Результаты разделения объектов класса на непересекающиеся группы предлагается оценивать по алгоритму в [2] использованием специальных мер компактности. Измерение компактности служат инструментом анализа изменений в структуре классов при удалении шумовых объектов.

Предлагаются оценку компактности структуры отношений объектов измерять в (0;1]. Существуют классические алгоритмы группировки объектов, в том числе алгоритм DBSCAN. Этот алгоритм кластеризации на основе плотности: для набора точек в заданном пространстве алгоритм группирует близко расположенные точки, отмечая как выбросы изолированные точки в области низкой плотности (включая дальних ближайших соседей). DBSCAN — один из наиболее часто используемых алгоритмов кластеризации. Если в пределах поискового расстояния от конкретной точки не найдено минимальное количество объектов кластера, то эта точка помечается как основная и включается в кластер вместе со всеми точками. В выборке [3] мы разделяем объекты на группы с помощью алгоритма DBSCAN и визуализируем их помощью алгоритма [5] . Алгоритм разделил выборку на две группы на 19 и 81 объектов.

Представление объектов в новом (двумерном) признаковом пространстве показано на рис. 1 и рис. 2. Мы можем визуализировать представление разных групп путем изменения значений радиусов в алгоритме DBSACAN. В табл. 1 приведены значения компактности для DBSCAN и связанности по системе пересекающихся интервалов.

Рис1. Визуализация разделения на две группы объектов в DBSCAN из 19 и 81 объектов

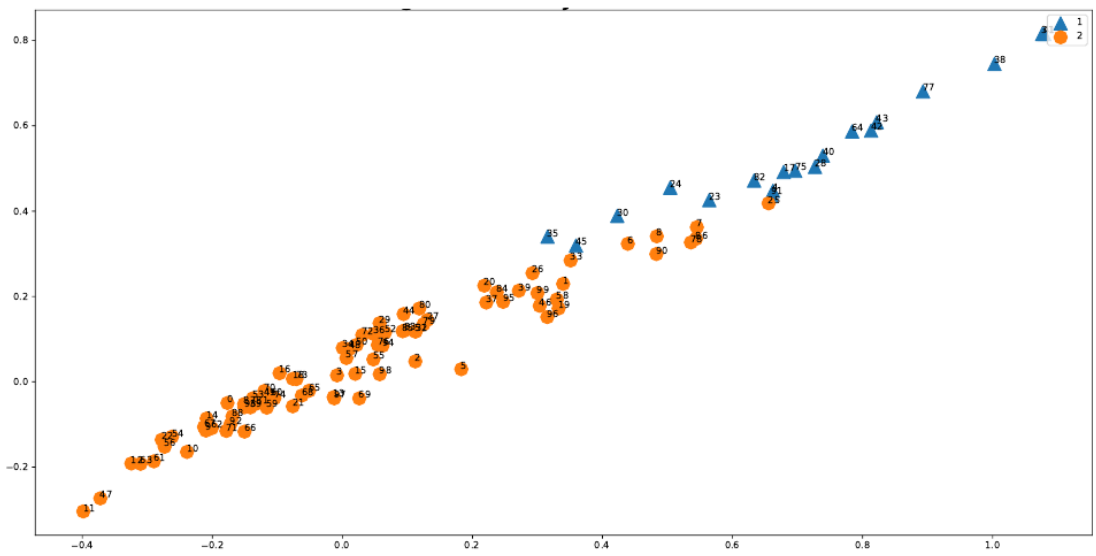
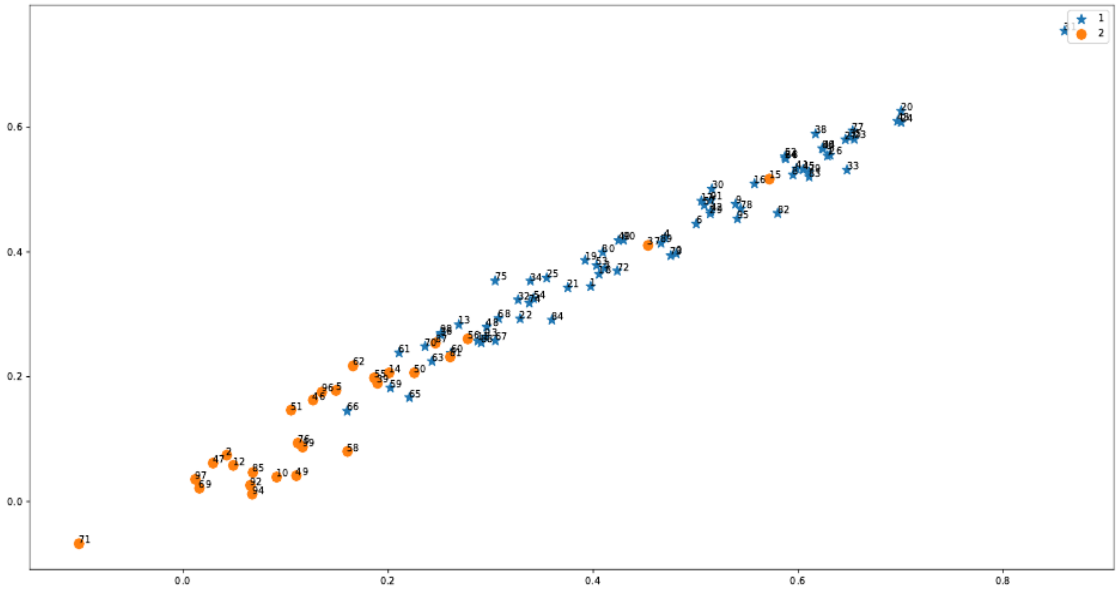


Рис2. Визуализация двух групп в DBSCAN в соотношении 72 и 28

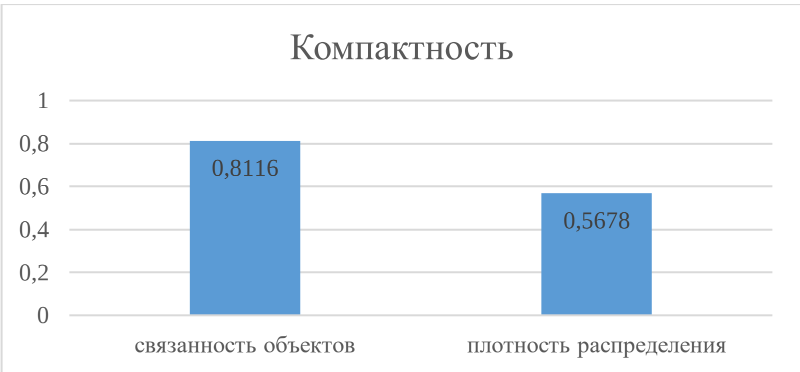


Мы можем сравнить эффективность использование двух отношений объектов по мере компактности (см. табл. 1).

Таблица 1. Сравнение отношений по мере компактности оценок

Отношение	Количество групп	Компактность
плотности	2	0.5678
связанности	4	0.8116

на графике ниже мы видим, что метод связанности объектов по пересекающимся гипершарам дает более высокую компактность, чем метод DBSCAN по плотности распределения.



Литература:

[1] Загоруйко Н.Г, Кутненко О.А . , Зырянов А . О. , Леванов Д.А Обучение распознаванию образов без переобучения // Машинное обучение и анализ данных, Т. 1 . – № 7. – 2014. – С . 89 1–90 1 .

[2] Ignatyev N. A . Structure Choice for Relations between Objects in Metric

Classification Algorithms // Pattern Recognition and Image Analysis, V. 28. – № 4. – 2018. – P. 590–597.

[3] Убайдуллаева Р.Т. Саморефлексия как субъектно-практическая методология социологии: // Автореферат диссертации доктора по социологическим наукам, Ташкент. –2019. – С. 11–28.

[4] Ignatyev N. A . On Nonlinear Transformations of Features Based on the Functions of Objects Belonging to Classes // Pattern Recognition and Image Analysis, V. 31 . – № 2. – 2021 . – P. 197–204.

[5] 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.

[6] Зиновьев А .Ю. Визуализация многомерных данных // Красноярск, Изд. КГТУ, – 2000. – 1 80 . с .