АНАЛИЗ СТРУКТУРЫ ОТНОШЕНИЙ ОБЪЕКТАМИ ВЫБОРКИ НА ОСНОВЕ МЕРЫ КОМПАКТНОСТИ

Мукумов Салим

Национальный университет Узбекистана

100174, Ташкент, Вузгородок 4,

E-mail: salimmuqumov2505@gmail.com

Анализ структуры отношений, определяемых расстояниями между объектами выборки данных, необходим для проверки ее (структуры) соответствия гипотезе о компактности классов в теории распознавания. Количественные оценки степени соответствия объектов этой гипотезе могут выражаться через меры компактности обучающей выборки. Как указывалось в [1], единой меры компактности не существует. Как правило, на значения оценок компактности влияют следующие факторы:

– выбор метрики для вычисления расстояний между объектами;

– значение размерности признакового пространства;

– выбор способа масштабирования и нормирования данных;

– использование методов отбора информативных наборов признаков;

– условия отбора и удаления шумовых объектов из выборки данных;

– количество объектов–эталонов минимального покрытия обучающей выборки;

– линейные и нелинейные преобразования признакового пространства для описания объектов.

Целью поиска экстремальных значений мер компактности на многообразии факторов из приведенного выше списка является повышение качества распознавания. На использовании меры расстояния основана реализация алгоритмов метода ближайший сосед (БС). Множество алгоритмов метода БС

имеет бесконечную ёмкость и в рамках классической теории Вапника–Червоненкиса нет возможности оценить их качество распознавания. Одним из способов обоснования этих алгоритмов является вычисление оценок обобщающей способности распознавания с помощью профиля компактности.

Общность между предлагаемой мерой компактности и обобщающей способностью алгоритмов заключается в определении и использовании кластерной структуры на множестве объектов обучающей выборки. Единственность её (меры) значений гарантируется методом разбиения объектов классов на непересекающиеся группы [2]. Разбиение основано на использовании определяемого подмножества граничных объектов (оболочек) классов по заданной (базовой) метрике и логических закономерностей в форме гипершаров. Потребность в использовании кластерной структуры с точки зрения качества обучения заключается в:

–обнаружении и удалении шумовых объектов;

–выделении объектов–эталонов минимального покрытия выборки без шумовых объектов, обеспечивающих корректное разделение ее на классы.

Порядок выполнения описанных выше действий необходим для того чтобы исключить отнесение шумовых объектов к множеству объектов–эталонов минимального покрытия выборки. Для вычисления качества обучения предлагается использовать специальный показатель, который выражается через среднее число объектов обучающей выборки без шумовых объектов, притягиваемых одним эталоном минимального покрытия. Показана связь этого показателя с показателями обобщающей способности алгоритмов распознавания прецедентного типа.

Последовательность реализации процедуры поиска минимального покрытия объектами–эталонами выборки E_{ob} определяется следующим образом. Упорядочим объекты каждой группы $G_u \cap K_t$, $u=1,\ldots,\delta$, $t=1,\ldots,l$ по множеству значений $\{R_S\}_{S\in G_u}$. В **качестве меры** расстояния между $S\in G_u$, и $=1,\ldots,\delta$ и произвольным допустимым объектом S' **используется локальная метрика**

$d(S,S')=\rho(S,S')/R_S$. Решение о принадлежности S' к **одному из** классов $K_1,\ldots,K_l$ или отказе от распознавания принимается по правилу: $S'\in K_t$ если

$$d(S_\mu,S')=\min_{S_j\in E_{ob}} d(S_j,S') \text{ и } S_\mu\in K_t \text{ и } d(S_\mu,S')\neq \min_{S_j\in CK_t\cap E_{ob}} d(S_j,S').$$

Согласно принципу последовательное исключение, используемого в процессе поиска покрытия, выборка E_{ob} делится на два **подмножества:** множество эталонов E_{ed} и контрольное множество E_k, $E_{ob}=E_{ed}\cup E_k$. В **начале** процесса $E_{ed}=E_{ob}$, $E_k=\varnothing$. Упорядочение по значениям отступа $\{R_S\}_{S\in G_u}$, и=1,…,δ 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 остаётся корректным (без ошибок распознающим объекты) на E_{ob}.

Будем считать, что нумерация групп объектов отражает порядок $|G_1|\geq \ldots \geq|G_\delta|$ и по группе G_p, p=1,…,δ 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 $S\in E_k$ **нарушает** корректность **решающего** правила **,** то S возвращается в множество E_{ed}.

Показателем компактности выборки E_0 при использовании правила является среднее число объектов E_{ob}, притягиваемых одним эталоном минимального покрытия из E_{ed}

$$\Omega(E_0,\rho)=\frac{|E_{ob}|}{|E_{ed}|}.$$

ФОЙДАЛАНИЛГАН АДАБИЁТЛАР РЎЙХАТИ: (REFERENCES)

1. Загоруйко Н.Г., Кутненко О.А., Зырянов А.О., Леванов Д.А. Обучение распознаванию

2. Игнатьев Н.А. Кластерный анализ данных и выбор объектов-эталонов в задачах распознавания с учителем // Вычислительные технологии, 2015. Т. 20. №6. С. 34–43.