

ПОСТРОЕНИЕ МИНИМАЛЬНОГО ПОКРЫТИЯ С ТЕСТАМИ ДЛЯ АЛГОРИТМОВ БЛИЖАЙШЕГО СОСЕДА

Кучкарова Маржонабону Каримжон кизи

*магистрант 1 курса Национального университета Узбекистана факультета
прикладной математики и интеллектуальных технологий в направлении
искусственный интеллект*

Аннотация. Рассматривается разбиение объектов на группы на основе отношений их связанности и плотности распределения. Плотность по локальной области в форме гипершара вычисляется относительно его центра в объекте выборки. Принадлежность объектов к непересекающимся классам определяется по объединению локальных областей с заданными ограничениями на значения плотности распределения. Определено условие выделения шумовых объектов классов из множества граничных. Проверка условия необходима для корректировки состава классов и группировки объектов по отношению их связанности по системе пересекающихся гипершаров. Обоснованы выделения эталонов минимального покрытия классов и их применение при оценке качества группировки. Для оценки формы конфигурации группы предложен показатель, значением которого является длина минимального кратчайшего пути между эталонами.

Ключевые слова: отношение связанности; оценки качества кластеризации; плотность распределения; шумовые объекты; эталоны минимального покрытия.

Введение

В задачах группировки с частичным обучением считается, что для части объектов выборки классификация уже известна. Классификация может быть задана экспертами либо введена по результатам разведочного анализа данных. Оценки плотности распределения в локальных областях признакового пространства рассматриваются в качестве примера такого анализа.

Для количественной оценки качества кластеризации используют внешние и внутренние меры [1– 3]. Внешние меры основаны на сравнении результата кластеризации с априори известным разделением на группы. Эти меры используют дополнительные знания о кластеризуемом множестве: распределение по кластерам, количество кластеров и т.д. Внутренние меры отображают качество кластеризации только по информации в данных.

Существующие меры оценки качества кластерного анализа не распространяются на результаты ряда алгоритмов. Если конфигурация кластеров не является сферической, то идея использования методов вычисления их центров является абсурдной. Центр может находиться за пределами конфигурации кластера. Например, вероятность такого события существует для результатов алгоритма кластеризации наборов

многомерных данных на основе плотности распределения DBSCAN [2]. Алгоритм может обнаруживать кластеры любой формы и автоматически определять выбросы.

Имеются исследования с описанием методов для понижения чувствительности результатов от выбора параметров ε и k для DBSCAN, в которых ε представляет максимальный радиус гипершара с центром в объекте наблюдения, а k – ограничение на минимальное количество объектов выборки, содержащихся в таком гипершаре [4]. Несмотря на все попытки внесения модификаций не было предложено ни одного критерия оценки качества кластеризации по алгоритму DBSCAN.

Для устойчивого разбиения на кластеры в [5] предложен метод локального осреднения для вычисления плотности распределения данных как точек в метрическом пространстве. При выделении областей

сгущения пространства применялось осреднение Гауссовского типа

$$P(x, x) = \exp \left(-\pi \sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right)$$

радиусу R , которое инвариантно относительно поворотов (отображений, не меняющих взаимные расстояния), и суммарная вероятность вхождения точки x_i в пространство равна 1. Обоснованы условия выбора значений локальных радиусов осреднения и радиусы могут различаться для разных кластеров.

DBSCAN может как разъединить кластер, так и объединить в один несколько различных кластеров. Существует потребность в разработке и использовании новых мер оценки качества кластеризации с учетом топологических свойств и плотности распределения объектов. В [6] проводились исследования топологических свойств по отношению связанности объектов классов по системе пересекающихся гипершаров. Условием связанности было наличие в пересечении граничных объектов классов. Шумовые объекты при машинном обучении рассматривались как подмножество граничных. Удаление шумовых объектов использовалось как средство цензурирования обучающих выборок с целью повышения обобщающей способности алгоритмов распознавания. Косвенной оценкой качества распознавания в порядковой шкале измерений служила мера компактности, значение которой зависело от

числа эталонов минимального покрытия выборки, шумовых объектов и их составов.

вхождения точек в шар единичного объема. Локаль-

Перечислены причины, по которым алгоритм

Разработка новых мер оценки качества кластерного анализа расширяет возможности обнаружения скрытых закономерностей в рамках информационных моделей, основанных на знаниях. В качестве нового знания экспертам из предметных областей становится доступной информация о структуре кластеров со сложной конфигурацией.

1. Предмет исследования

При отсутствии каких-либо предположений о природе среды для анализа данных в локальных областях признакового пространства используют непараметрические методы оценки плотности распределения. В методах « k ближайших соседей» и «парзеновское окно» локальные области для вычисления плотности имеют сферическую форму. Геометрическое разнообразие форм конфигурации локальных областей с высокой плотностью распределения свойственно для результатов группировки по алгоритму DBSCAN [4]. Потребность в анализе разнообразия форм конфигураций послужило причиной разработки новых критериев (индексов) качества кластеризации на основе топологических свойств объектов классов.

Для анализа качества кластеризации с учетом топологии признакового пространства необходимо иметь данные о принадлежности хотя бы части объектов выборки к непересекающимся классам. Включение таких данных об объектах при кластерном анализе рассматривается как задание частичного обучения. Процесс выделения множества шумовых объектов классов интерпретируется в новом смысле. Шумовой объект – это кандидат на смену значения целевого признака класса или перехода из состава одной группы в другую. Представление о форме конфигурации группы формируется по результатам анализа структуры отношений между эталонами минимального покрытия классов.

Введение классификации на множестве не разделенных по категориям объектов рассматривается как предобработка данных с использованием эвристик. Выбор эвристик может быть связан с формой представления локальной области и вычислением значений плотности распределения в ней. Например, с представлением в виде гипершаров с фиксированным радиусом и центрами в каждом объекте выборки, вычисление значений плотности распределения в которых зависит как от числа объектов, так и от их расстояний от центра.

Классификация объектов на основе плотности их распределения связана с использованием метрик, методов нормирования данных, выбором размеров локальной области, правилами деления множества объектов на непересекающиеся подмножества. Такие термины и понятия, как граничный объект, меры компактности, эталоны минимального покрытия, информативные признаки, шумовые объекты, гипотеза о компактности, используемые при классификации, становятся востребованными и при оценке качества кластерного анализа.

Следующая последовательность операций над данными из табличного набора (выборки объектов) служит примером предобработки для введения классификации:

- вычисление плотности распределения объектов выборки;
- формирование упорядоченной по значениям плотности последовательности;
- разбиение последовательности на непересекающиеся подпоследовательности (классы). Наличие классификации объектов позволяет:
- выделять шумовые объекты классов по множеству граничных;
- корректировать составы классов за счет шумовых объектов;

- производить разбиение объектов классов на непересекающиеся группы по отношению их связанности;
- определять эталоны минимального покрытия;
- вычислять принадлежность произвольного допустимого объекта к группе по мере его близости к эталонам минимального покрытия.

Условия принадлежности граничного объекта к множеству шумовых при кластерном анализе

оставались неисследованными. Разъясним смысл, придаваемый понятиям шумового и граничного объектов при классификации, и покажем, как их можно адаптировать для кластерного анализа.

Разбиение объектов на непересекающиеся группы [6] с учетом топологических свойств признакового пространства производится в каждом классе по системе пересекающихся гипершаров. Количество групп вычисляется по отношению связанности, и их состав зависит от множества граничных объектов. Частичное представление о форме конфигурации группы дают эталоны минимального покрытия. Структура расположения эталонов зависит от множества граничных по заданной метрике объектов классов. Изменение состава эталонов и структуры отношений между ними связано с выделением шумовых объектов из множества граничных. Перенос шумовых объектов из одного класса в другой является причиной изменения числа и состава эталонов на выборке данных.

При включении граничного объекта [6] в состав шумовых использовались отношения расстояний до ближайших объектов из своего класса и его дополнения. Для получения оптимального с точки зрения точности распознавания решения вводились ограничения (пороги) на значения оценок отношений. Удаление шумовых объектов как части граничных в обучающей выборке приводило к увеличению отступа между классами и рассматривалось в качестве средства повышения обобщающей способности алгоритмов распознавания. Отступ является относительной величиной, которая влияет на вычисление значений параметров локальных метрик и состав эталонов минимального покрытия выборки.

В отличие от задач классификации при кластерном анализе шумовой объект не удаляется из выборки. Условия на отступы между граничными объектами классов определяют переход из одной группы в другую.

Разработка новых внутренних критериев для кластерного анализа и получения дополнительных сведений о форме конфигурации групп связана с условиями вычисления отступов между классами. Предлагается перечень дополнительных сведений (знаний) о конфигурации группы, которые можно получить по результатам кластерного анализа:

1. Число и описание эталонов группы.
2. Длина кратчайшего незамкнутого пути (КНП) между эталонами.

3. Число эталонов, являющихся граничными объектами группы.
 4. Среднее число объектов группы, притягиваемое одним эталоном минимального покрытия.

2. Постановка задачи. Классификация объектов на основе плотности их распределения

Для описания множества (выборки) объектов $E0 = \{S1, ..., Sm\}$ применяется набор разнотипных признаков $X(n) = (x1, ..., xn)$. Определены процедуры для:

– вычисления плотности распределения $\beta_i = \phi(Si, \varepsilon)$ объекта $Si \in E0$ в локальной области в форме гипершара радиуса ε по метрике $\rho(x, y)$;

– деления объектов $E0$ по множеству $\{\beta_i\}$ на непересекающиеся классы $K1, K2$. Объекты $Su, Sv \in Ki, i = 1, 2$, считаются связанными между собой (обозначается $Su \leftrightarrow Sv$), если

$\{S \in B(E0, \rho) | \rho(S, Su) < ru \text{ and } \rho(S, Sv) < rv\} \neq \emptyset$, (1) где ru (rv) – расстояние до ближайшего от Su (Sv) объекта из $K3-i$ по метрике $\rho(x, y)$, $B(E0, \rho) =$

$= \{S \in E0 | \rho(Su, S) = \min_{u=1, ..., m} \rho(Su, Sv)\}$ – множество граничных объектов классов, опреде-

$\{Su \in Kq, Sv \in K3-q$

ляемое

на

$E0$.

Множество $B(E0, \rho)$ используется для описания объектов $E0$ в пространстве из бинарных при-

знаков по таблице $T = \{tij\}_{m \cdot d}$, $d = |B(E0, \rho)|$. Значение элемента таблицы $tij = 1$, если выполняется условие (1) и $tij = 0$ в противном случае. По таблице T определяется минимальное число групп из связанных и несвязанных объектов классов. Составы групп не пересекаются и представлены объектами класса $K1$ или $K2$. Для любых объектов $Sb, Sa \in Ki$, принадлежащих одной группе, всегда существует множество пар пересекающихся по $B(E0, \rho)$ гипершаров для формирования цепочки $Sa \leftrightarrow St \leftrightarrow ... \leftrightarrow Sb$.

Требуется:

– произвести разбиения объектов $E0$ в каждом классе $Kt, t = 1, 2$, по отношению связности (1) на непересекающиеся группы $Jt1, ..., Jtl(t)$ и определить объекты минимального покрытия эталонами;

69

Обработка информации / Data processing

– оценить качество разбиения на группы по классам и выборке в целом.

Пусть размер локальной области задан по значению радиуса гипершара ε . Значение плотности распределения по гипершару с центром в объекте $S \in E0$ определим как

$\phi(S, \varepsilon) = \sum_{S \in B(E0, \rho)} 1 -$

$\rho(S, S)$

ε

$\rho(S, Si) < \varepsilon$

$i \mid (2) \setminus$

Тогда значением математического ожидания по (2) будет

$$m M(\varepsilon) = \sum \varphi(S, \varepsilon) / m.$$

По (2), (3) определим исходный (без корректировки) состав классов: $K1 = \{Sj \mid \rho(Sk, Sj) < \varepsilon \text{ and } \varphi(Sk, \varepsilon) > M(\varepsilon)\}$ и $K2 = E0 \setminus K1$. (4)

Корректировка состава классов по (4) связана с отбором шумовых объектов. Число групп по отношению связанности определяется алгоритмическим путем.

С изменением структуры отношений внутри групп и между группами связан процесс выделения шумовых объектов и перехода из состава одной группы в другую. В рамках поставленной в работе задачи результаты отнесения объекта к множеству шумовых зависят от значения параметра ε для вычисления плотности распределения по (2) и классификации по (4).

Выделение шумовых объектов по (5) используется для коррекции состава множества граничных объектов классов. Коррекция граничных объектов приводит к изменению отношений связанности объектов при разбиения их на группы. Меняются число и состав эталонов минимального покрытия. Исследовать структуру отношений внутри группы можно через КНП между эталонами минимального покрытия и расстояния объектов до них. Анализ расстояний до эталонов и их интерпретация для оценки качества кластерного анализа в данной работе не рассматриваются. Значения КНП вычисляются при числе эталонов больше или равно 2. Источниками информации для обнаружения скрытых закономерностей по конфигурации группы являются:

- пересечение множества эталонов и граничных объектов;
- соотношение между числом эталонов и числом граничных объектов;
- среднее число объектов группы, притягиваемое одним эталоном минимального покрытия или

компактность

группы.

По результатам анализа эталонов и граничных объектов из групп можно получить представле-

ние об их конфигурациях. Например, как это показано на рис. 2.

Рис. 2. Визуальное представление о конфигурациях групп Fig. 2. Visual representation of group configurations

Визуальное представление на рис. 2 группы из класса $K2$ можно интерпретировать следующим образом:

- все эталоны являются граничными объектами;
- среднее число объектов группы, притягиваемое одним эталоном минимального покрытия, меньше, чем у группы из $K1$.

4. Внешняя мера оценка качества кластеризации

Пусть $J11, \dots, J1u$ ($J11 \cup \dots \cup J1u = K1$) и $J21, \dots, J2v$ ($J21 \cup \dots \cup J2v = K2$) – множества непересекающихся групп, полученных по отношению связанности объектов (1) по классам $K1$ и $K2$. Каждая группа определяется по непересекающемуся

подмножеству граничных объектов $B(E0, \rho)$, принадлежащих одному из классов: $K1$ или $K2$. Мера компактности кластеров по классам и выборки в целом [7] используется для оценки качества кластеризации и определяется как

Заключение

В статье предложена новая методика оценки качества кластеризации при интеллектуальном анализе данных групп объектов со сложной конфигурацией. Методика ориентирована на исследование топологических свойств объектов через их классификацию в локальных областях многомерных данных с различными плотностями распределений. Численная оценка топологических свойств выражается через значения меры компактности групп по отношению связности объектов классов. Отношение связности применялось для поиска минимального покрытия выборки эталонами. Длина КНП между эталонами группы служит информацией для формирования представления о ее (группе) конфигурации.

СПИСОК ИСТОЧНИКОВ:

1. Kleinberg J. An Impossibility Theorem for Clustering. URL: <https://www.cs.cornell.edu/home/kleinber/nips15.pdf> (accessed: 02.06.2024).
2. Сивоголовко Е.В. Методы оценки качества четкой кластеризации // Компьютерные инструменты в образовании. 2011. No 4. С. 14–31.
3. Шмидт Ю.Д., Лободина О.Н., Грищенко В.В., Нинбо Ц. Об оценке уровня инновационного развития регионов России // Региональная экономика: теория и практика. 2017. Т. 15, No 2. С. 342–354.
4. Shili H. Clustering in big data analytics: a systematic review and comparative analysis (review article)