Выбор целевых признаков для классификации и кластерного анализа структур отношений объектов

Н. А. Игнатьев

Аннотация **– Рассматривается проблема анализа качества кластеризации объектов с произвольной (несферической) формой конфигурации. Используется присвоение категорий (высокая, низкая) объектам на основе плотности их распределения в локальных областях признакового пространства. Определяется подмножество граничных объектов по отношению к объектам другой категории. Вводится понятие метаобъекта для случая многомодального распределения данных. Для описания метаобъекта используются значения устойчивости признаков для объектов из двух классов, соответствующих разным модальностям. Значения устойчивости по множеству пар классов формируют унифицированное пространство из новых признаков для описания метаобъектов. Общие ограничения на множество их допустимых значений объясняются свойствами функций принадлежности в теории нечёткой логики. Для анализа качества кластеризации метаобъектов применяется их деление на категории по плотности распределения. Единственность числа и состава групп с несферической формой конфигурации обеспечивается за счёт использования свойств отношений связанности объектов одной категории по системе пересекающихся гипершаров. Центрами гипершаров являются объекты групп. В пересечении двух и более гипершаров находится хотя бы один граничный объект. Для вычисления мер внутригрупповой близости и межгруппового различия используются расстояния до эталонов минимального покрытия в составе группы и подмножества граничных объектов. Оценки качества кластеризации для каждой группы выводятся на основе соотношения этих мер. Предлагается условие для значений оценок, определяющее компактность группы. Предложен метод оценки качества кластеризации объектов с произвольной формой конфигурации с учётом плотности распределения и топологии их размещения. Топологические свойства выражаются через отношение связанности объектов в зависимости от категории их плотности. На реальных данных демонстрируется их отображение в унифицированное признаковое пространство с последующей оценкой качества кластерного анализа метаобъектов. Технология формирования унифицированного пространства признаков для описания метаобъектов может быть полезной в задачах с пропущенными данными и при наличии многомодального распределения.**

Ключевые слова **– оценки качества кластеризации, эталоны минимального покрытия, устойчивость признаков, унифицированное признаковое пространство**

Статья получена 10 декабря 2025.

Н.А. Игнатьев – Национальный университет Узбекистана, Ташкент, Узбекистан (e-mail: n_ignatev@rambler.ru),

I. Введение

Кластерный анализ данных играет важную роль для построения моделей в слабо формализованных предметных областях. Многообразие структур отношений объектов (признаков) является предметом исследования при обосновании выбора методов для таких моделей. Существенными недостатками методов кластерного анализа являются:

– отсутствие единого критерия качества кластеризации. По этой причине для определения качества и оценки состава выделенных кластеров требуется привлечение эксперта предметной области;
– реальное число кластеров, как правило, заранее не известно и выбирается по субъективным соображениям.

Наличие или отсутствие классификации предопределяет выбор методов интеллектуального анализа данных для решения задач в информационных моделях, основанных на знаниях. Результаты группировки или вычисления плотности распределения могут использоваться для определения категорий объектов на выборках данных. При реализации ряда методов группировки объектов и признаков изначально предполагается наличие разбиения объектов на классы.

Существуют примеры решения проблем Big Data комбинированием методов классификации и кластерного анализа. Разбиение объектов на классы позволяет:

– производить редукцию пространства через группировку разнотипных признаков и формирование из них латентных показателей методом обобщённых оценок [1];
– использовать алгоритмы кластерного анализа в латентном признаковом пространстве при отборе объектов в базу прецедентов для машинного обучения алгоритмов с целью распознавания DDoS атак [2].

Для обоснования результатов разбиения объектов на группы алгоритмами кластерного анализа были введены специальные критерии и аксиомы. Доказана теорема невозможности Клейнберга [3] об отсутствии оптимального алгоритма кластеризации. Выводы основываются на трёх аксиомах (о масштабной инвариантности, полноте и согласованности), которые одновременно не могут выполнятся при кластерном анализе.

Различают внешние и внутренние меры для количественной оценки качества кластеризации. Внешние 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.

Особенности реализации ряда алгоритмов не позволяют интерпретировать их результаты с помощью существующих мер оценки качества кластерного анализа. К числу таковых особенностей относятся:

– кластеры имеют сложную и изначально неопределяемую форму конфигурации;

– существуют выбросы в виде объектов выборки, принадлежность к группам которых не определена.

Исследование отношений объектов классов в [4] основывается на гипотезе о компактности. Разбиение на группы объектов по каждому непересекающемуся классу рассматривалось как предобработка данных для поиска эталонов минимального покрытия выборки. Число групп и их состав определялись алгоритмическим путём. Для вычисления двух мер компактности как количественных показателей отношений объектов классов использовались мощность групп и среднее число объектов, притягиваемых одним эталоном минимального покрытия. Показано, что эти показатели [5] имеет смысл применять для вычисления внешней и внутренней мер при оценке качества кластерного анализа.

II. Предмет исследования

На вычислении значений плотности распределений по эмпирическим данным в локальных областях признакового пространства основана реализация ряда алгоритмов кластерного анализа. К числу таковых относятся методы, в которых не используются предположения о виде плотности распределений данных. Например, для кластеров, определяемых по алгоритму DBSCAN [6]. Плотность распределения этим алгоритмом оценивается по двум заданным параметрам: k – количеству ближайших соседей и ε – радиусу гипершара. Статусы объектов (достижимый, граничный, выброс), определяющих результаты группировки, не являются постоянными при разных значениях k и ε. Форма конфигурации кластеров изначально неизвестна.

Актуальной проблемой является исследование устойчивости к выбросам (робастности) методов в условиях, когда плотность распределения, как правило, неизвестна. Существующие внешние и внутренние меры оценки качества кластеризации для этих целей не всегда подходят. Например, из-за особенностей реализации алгоритма DBSCAN нет возможности для вычисления центроидов групп и учёта ограничений на шкалы измерений признаков [7]. Соблюдение баланса между значениями k и ε затруднено, поскольку реальные выборки данных разбиваются на группы с разной плотностью и границами разной степени размытости.

Значения плотности распределения в локальных областях признакового пространства находят применение для разбиения объектов на непересекающиеся классы. Вычисление статуса объекта достижимый, граничный и выброс по параметрам k и ε алгоритмом DBSCAN может рассматриваться в качестве информации для задания значения целевого признака при классификации. Для аналогичных целей также востребованы результаты анализа мультимодального распределения данных.

По значению ограничения (порога) на плотность распределения объектов в гипершарах заданного радиуса ε в [5] производилось их разбиение на классы. Целью разбиения было оценить структуру отношений объектов классов по группам с относительно высокой и низкой плотностью распределений. Определены условия выбора кандидатов на смену составов групп при формировании множества граничных объектов классов. В [5] для оценки групп со сложной конфигурацией использовалась длина кратчайшего незамкнутого пути между эталонами минимального покрытия. Анализ сложности конфигурации производился для групп с числом эталонов больше или равном 2.

При классификации граничные объекты рассматриваются как ближайшие объекты из противоположных классов, посредством которых вычисляются выбросы (шумовые объекты), параметры решающих функций и т. д. Граничные объекты классов используются для оценки топологической структуры (не природы среды) отношений объектов в признаковом пространстве. Для оценки используются значения меры компактности объектов [4] как для отдельных классов, так и для всей выборки. При вычислении меры компактности используются множество гипершаров с центрами в объектах классов. Непустые пересечения гипершаров содержат хотя бы один граничный объект одного с их центрами класса.

Определены условия формирования множества шумовых объектов из множества граничных объектов классов [5]. Принадлежность объектов к множеству шумовых используется для переноса их из состава одной группы в другую. Смена состава групп изменяет конфигурацию граничных объектов классов. Изменяется размещение и состав эталонов минимального покрытия. В качестве значения внутренней меры предлагается использовать соотношение внутрикластерного сходства и межкластерного различия, вычисляемое по множеству эталонов минимального покрытия. Неисследованной остаётся зависимость межкластерного различия от эталонов минимального покрытия, полученных по базовой или локальным метрикам.

В алгоритме метода обучения распознаванию [4] определено условие, согласно которого представители из разных классов не должны входить в одни группы, сформированные по отношению связанности объектов по системе пересекающихся гипершаров. Число групп и постоянство их состава при наличии такого условия определяется алгоритмическим путём. По отношению связанности объектов классов форма конфигурации групп может быть различной. Каждая группа идентифицируется как минимум одним эталоном. Нет необходимости в вычислении центров групп, во введении ограничений на шкалы измерений признаков.

Многообразие разбиений объектов на два класса с высокой плотностью K_1 и относительно низкой плотностью распределения K_2 ограничено и зависит от числа k ближайших соседей или объёма локальной области по радиусу ε по отдельности или от их совместного использования в алгоритме DBSCAN. Эту

зависимость предлагается исследовать для группировки по отношению связанности объектов через новые внутренние меры оценки качества кластеризации. Для анализа групп с произвольной формой конфигурации рассматриваются эталоны минимального покрытия. Представление о форме конфигурации дают результаты анализа пересечения множества эталонов покрытия и граничных объектов классов.

Проблема поиска уникального числа тем при решении задач тематического моделирования обсуждалась в [8]. Для проверки гипотезы о существовании уникального числа в рамках данного исследования предлагается присваивать категории объектам(документам). К классу K_1 следует относить документы из групп с семантически связанным контентом, в K_2 – документы из групп, интерпретируемых как фоновые.

Применение новых мер оценки качества кластеризации позволяет делать выводы об эффективности использования для группировки различных метрик, способов нормирования данных, отбора наборов признаков и т.д. Эти выводы рассматриваются как источник нового знания для экспертов из предметных областей, получаемых ими при проверке своих гипотез на данных через решение задачи кластерного анализа из ограниченного числа многообразий группировок объектов.

Сложность обнаружения скрытых закономерностей из данных часто связывают с проблемами Big Data. Редукция признакового пространства с помощью метода главных компонент (PCA) является частным примером решения проблемы проклятия размерности в Big Data при машинном обучении. Сокращение числа объектов, разделённых по категориям, рассматривается как процесс поиска эвристик для решения задачи минимального покрытия выборок данных эталонами. Необходимы критерии для сравнения и обоснования эвристик, интерпретация полученных с их помощью результатов.

Наличие категорий у объектов позволяет [1] отображать их описания из исходного пространства в унифицированное с однозначно интерпретируемыми свойствами признаков. Метод отображения основывается на применении функций принадлежности к нечётким множествам. Элементы нечётких множеств представлены значениями признаков объектов классов. В унифицированном пространстве есть возможность проводить кластерный анализ по категориям объектов.

Существует потребность в проведении исследований связанных с:

– определением категорий объектов выборок данных по их статистическим и топологическим параметрам;

– оценкой топологии групп с учётом граничных объектов и эталонов минимального покрытия классов.

III. Математическая модель

В множестве (выборке) $E_0=\{S_1,\ldots,S_m\}$ каждый объект $S_i \in E_0$ описывается набором разнотипных признаков $X(n)=(x_1,\ldots,x_n)$. На множестве E_0 определена процедура вычисления плотности распределения объектов по метрике $\rho(x,y)$ и задано их разбиение на два непересекающихся класса K_1 и K_2. Считается, что в K_1 включены объекты с относительно высокой плотностью, в K_2 – низкой, на которых определяется отношение связанности объектов по системе пересекающихся гипершаров. Объекты $S_i, S_j \in K_q$, $q=1,2$ считаются связанными между собой ($S_i \leftrightarrow S_j$), если

$$\{S \in B(E_0,\rho) | \rho(S,S_i) < r_i \ \ and \ \rho(S,S_j) < r_j\} \neq \varnothing, \quad (1)$$

где $r_i(r_j)$ – расстояние до ближайшего от $S_i(S_j)$ объекта из K_{3-q} по метрике $\rho(x,y)$,

$$B(E_0,\rho) = \left\{ S \in E_0 \,\middle|\, \rho(S_u,S) = \min_{S_u \in K_q, S_v \in K_{3-q}} \rho(S_u,S_v), u = 1,\ldots,m \right\}$$

– множество граничных объектов классов, определяемое на E_0.

Множество $G_{qv}=\{ S_{v_1},\ldots,S_{v_c} \}$, $c \geq 2$, $G_{qv} \subset K_q$, $q=1,2$, $v \leq |K_q|$ представляет область (группу) со связанными объектами в классе K_q, если для любых $S_{v_i}, S_{v_j} \in G_{qv}$ существует путь $S_{v_i} \leftrightarrow S_{v_k} \leftrightarrow \ldots \leftrightarrow S_{v_j}$.

Объект $S_i \in Kq$ 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 $S_j \in K_q$.

Множество $B(E_0,\rho)$ используется для описания объектов E_0 в пространстве из бинарных признаков по таблице $T=\{t_{ij}\}_{m \cdot d}$, $d=|B(E_0,\rho)|$. Значение элемента таблицы $t_{ij}=1$, если выполняется условие (1) и $t_{ij}=0$ в противном случае. По таблице T определяется минимальное число групп из связанных и несвязанных объектов классов. Составы групп не пересекаются и представлены объектами класса K_1 или K_2.

Считается, что на $K_1 \cup K_2$ определены процедуры:

– отбора минимального покрытия эталонами $E_{об} \subset E_0$ по группам по базовой и локальным метрикам [4];

– вычисления значений внутригруппового сходства и межгруппового различия.

Требуется разработать меры оценки качества кластеризации с учётом произвольной(несферической) конфигурации групп по:

– значениям внутригруппового сходства и межгруппового различия;

– множеству граничных объектов и эталонов минимального покрытия классов.

О многообразиях разбиения объектов на классы по значениям их плотности распределения. Использование мер расстояния, способов нормирования, методов отбора информативных признаков или выбора латентного признакового пространства изменяет структуру отношений близости объектов и их плотности распределения. Математическое ожидание или медианное значение плотности, вычисляемое на выборке данных по локальным областям из k ближайших соседей или по радиусу гипершара ε, являются параметрами для разбиения объектов на классы.

В качестве способа уточнения результатов разбиения в [5] предложено использовать отбор шумовых объектов как кандидатов на переход из одного класса в другой. Условие корректировки составов классов основывалось на анализе значений относительного отступа между граничными объектами классов. Корректировка составов приводила к изменению отношений

связанности объектов классов (1) и результатов группировки на их основе.

Определение статусов объектов достижимый, граничный и выброс в процессе реализации алгоритма DBSCAN рассматривается как предобработка данных для введения классификации. Рекомендуемый вариант разделения на два класса: к K_1 относится объекты со статусами достижимый и граничный, к K_2 – выбросы. При таком варианте есть проблемы в регулировании соотношений между числом объектов из K_1 и K_2 через подбор значений k и ε.

Оценки качества кластеризации для групп с несферической формой конфигурации

Форма конфигурации групп по алгоритму DBSCAN и по отношению связанности объектов классов не является фиксированной, допускаются описания объектов признаками из разных шкал измерений. Центры групп могут находится за границами их конфигураций. Эталоны минимального покрытия $E_{об}$ являются подмножеством объектов выборки E_0. Пересечение $E_{об}$ с множеством объектов группы даёт косвенное представление о её конфигурации. Для оценки качества кластеризации для групп с несферической конфигурацией предлагается использовать эталоны минимального покрытия классов K_1 и K_2.

Состав эталонов из $E_{об}$ и их количество зависят от выбора базовой или локальных метрик для алгоритма, реализующего жадную стратегию отбора минимального покрытия методом последовательного исключения по каждой группе из $K_1=G_{11}\cup\ldots\cup G_{1u}$, $u\geq 1$ и $K_2=G_{21}\cup\ldots\cup G_{2v}$, $v\geq 1$. Локальная метрика для объекта $S\in K_t$, $t=1,2$ определяется по базовой $\rho(x,y)$ как $\rho_s(x,y)=\omega_s\cdot\rho(x,y)$, где $\omega_s=1/\rho(S,S^*)$, $\rho(S,S^*)=\min\limits_{S_r\in K_{3-t}}\rho(S,S_r)$. Число эталонов по локальным метрикам, как правило, меньше или равно аналогичного числа по базовой метрике. Данное утверждение легко проверяется по результатам вычислительного эксперимента.

Для удобства идентификации предлагается использовать индексацию групп без указания принадлежности к классам входящих в них объектов. Через G_j будем обозначать группу из связанных по (1) объектов и M_j – множество эталонов минимального покрытия из G_j, $M_j\subset G_j$, $|M_j|\geq 1$, $\bigcup\limits_{i=1}^{d} M_i = E_{ob}$, $d\geq 2$.

Способы для оценки отношений объектов в группах и между группами.

При реализации ряда методов кластеризации необходимо задавать число групп, которое, как правило, является свободным параметром. Для обоснования значений параметра используют заключения экспертов или специальные критерии, вычисляемые по мерам близости объектов в группах и между группами. Одним из способов для оценки качества кластеризации со множеством допустимых значений в [-1;1] является метод силуэтов [9]. Метод позволяет оценивать, насколько хорошо объекты из разных групп различаются друг от друга с учётом взаимной близости внутри своих групп.

Предлагается разработка новых методов для анализа качества кластеризации с учётом произвольной конфигурации групп, сформированных по отношению связанности объектов (1). При определении отношений связанности и группировки на её основе используется множество граничных объектов классов. Косвенным показателем сложности конфигурации группы $G_j\subset E_0$, $j\geq 2$ служит непустое пересечение множества её объектов с множеством эталонов минимального покрытия $E_{об}$, $M_j=G_j\cap E_{об}\neq\varnothing$.

Насколько хорошо объекты групп различаются друг от друга оценивается по расстояниям до эталонов и граничных объектов. Как минимум один граничный объект входит в состав группы, формируемой по отношению связанности объектов классов. Введём следующие обозначения расстояний от объекта $S\in K_t$, $t=1,2$ до:

– граничного объекта из K_{3-t} $a(S)=\min\limits_{S^*\in K_{3-t}}\rho(S^*,S)$;

– эталона минимального покрытия из $K_{3-t}\cap E_{об}$ $b(S)=\min\limits_{S^*\in K_{3-t}\cap E_{об}}\rho(S^*,S)$;

– эталона минимального покрытия из $K_t\cap E_{об}$ $c(S)=\min\limits_{S^*\in K_t\cap E_{об}}\rho(S^*,S)$.

По расстояниям $a(S)$, $b(S)$, $c(S)$ вычисляются оценки

$$\lambda(S)=\frac{a(S)}{2b(S)} \quad (2)$$

со множеством допустимых значений [z;0,5],z<0,5 и

$$\omega(S)=\frac{b(S)-c(S)}{b(S)} \quad (3)$$

соответственно со значениями из [z;1], z≤0.

Отметим выводы, которые следуют из анализа оценок (2) и (3) для объектов выборки данных E_0. Для $S\in K_t$, $t=1,2$ при $\lambda(S)=0.5$ ближайший граничный объект из K_{3-t} является эталоном минимального покрытия. Если

$$|G_r|=|\{S\in G_r|\omega(S)>0\}|, \quad (4)$$

то группа G_r является компактной. Считается, что на выборке E_0 определено разбиение объектов на компактные группы с произвольной формой конфигурации по условию (4), если $|\{S\in E_0|\omega(S)>0\}|=m$. Очевидно, что $E_{об}=\{S\in E_0|\omega(S)=1\}$.

Анализ структуры отношений объектов может быть дополнен значением длины кратчайшего незамкнутого пути (КНП) между эталонами минимального покрытия M_j группы G_j при $|M_j|\geq 2$.

О предобработке данных для группировки объектов.

Потребность в предобработке данных для реализации методов машинного обучения связана с решением следующих проблем:

– проклятием размерности или размыванием отношений объектов при использовании мер расстояния;

– ограниченностью ресурсов из-за комбинаторной сложности алгоритмов при обработке больших объёмов данных.

Присвоение категорий (меток классов) объектам позволяет производить редукцию пространства методами отбора и выбора признаков. Метками классов могут служить временные интервалы, свойства

объектов, определяемые по результатам экспертизы. Например, временным интервалом в медицине является возраст пациента, в математической лингвистике месяц или год издания документов.

При наличии классификации объектов есть возможность реализовать:

– преобразования данных, инвариантных к масштабам измерений признаков;

– отображение данных из описаний в номинальной и интервальных шкалах измерений в унифицированное признаковое пространство;

– формирование наборов латентных признаков линейными и нелинейными методами вычисления обобщённых оценок [1];

– редукцию пространства через удаление малоинформативных признаков.

Унифицированное пространство формируется из значений устойчивости признаков объектов [1], разделённых на два непересекающихся класса. Например, в задаче по информационной безопасности в паре классов (K_0,K_i) объекты K_i представляют описания сетевых протоколов по i-му типу DDoS атак, K_0 – нормального трафика. Для исследования структуры связей типов DDoS атак в унифицированном пространстве можно использовать методы кластерного анализа.

В процессе унификации значениям признака в описании объектов двух классов ставится в соответствие показатель его устойчивости из интервала (0;1]. Множество допустимых значений из (0;1] формируются по нелинейным преобразованиям данных с использованием функций принадлежности к двум классам. Устойчивость признака является комбинаторной оценкой, вычисляемой на реальных выборках данных.

Показателем информативности признака служит близость значения его устойчивости к 1.0. Для редукции пространства необходимо использовать упорядоченную по устойчивости признаков последовательность. Фундаментальным свойством устойчивости является доказанная сходимость по вероятности к фиксированному значению на выборках из генеральной совокупности [1].

Векторы из значений устойчивости признаков, полученные на различных комбинациях из пар классов, считаются представлениями метаобъектов в унифицированном пространстве. Выбор комбинаций зависит от целей исследования. В медицине описание по набору симптомов и синдромов множества пациентов определённого возраста может рассматриваться как отдельный класс. Метаобъект по двум возрастным категориям представляется вектором, размерность которого равна числу симптомов и синдромов. Из-за малых вариаций на выборках из генеральной совокупности значения устойчивости можно вычислять и при наличии пропусков (неизмеренных значений признаков) по медицинским данным у части пациентов. Примером множества пар классов $\{(K_0,K_i)\}$ для формирования набора метаобъектов служит выбор в качестве K_0 пациентов младшего или старшего возраста, а в качестве K_i пациентов отличных от них возраста. Допускается формирование класса K_0 из абстрактных объектов с описанием по множеству допустимых значений признаков.

IV. Вычислительный эксперимент

Для эксперимента использовалась данные Wine Quality [10]. Множество оценок качества белого вина по выборке E_0 из 4898 проб(объектов) по набору $X(11)=(x_1,\ldots,x_{11})$ количественных признаков представлены значениями из $\{3,\ldots,9\}$. Подмножество объектов с одинаковой оценкой качества, считалось отдельным классом.

Процесс формирования вектора из значений признаков в унифицированном пространстве объясним на примере двух классов K_1 и K_2, $h=|K_1 \cup K_2|$. Пусть для количественного признака $x_c \in X(11)$ в описании объектов $K_1 \cup K_2 \subset E_0$ построена упорядоченная по неубыванию последовательность

$$r_1,\ldots,r_v,\ldots,r_h. \quad (5)$$

При разбиении (5) на непересекающиеся интервалы их число p_c, $(p_c \geq 2)$ считается неизвестным. Согласно условию разбиения в границах каждого интервала частота встречаемости значений признака из описаний объектов класса K_t больше чем в K_{3-t}, $t=1,2$. Это условие используется в критерии для поиска минимального покрытия (5) множеством из p_c $(p_c \geq 2)$ непересекающихся интервалов $\{[r_u;r_v]^i\}$, $1 \leq u$, $u \leq v \leq h$, $i=1,\ldots,p_c$.

Пусть $d_{tc}(u,v)$, $d_{3-t,c}(u,v)$ – количество представителей классов K_t, K_{3-t} в интервале $[r_u;r_v]^i$, $i \in \{1,\ldots,p_c\}$. В рекурсивной процедуре выбора значений r_u, r_v используется критерий

$$\left| \frac{d_{tc}(u,v)}{|K_t|} - \frac{d_{3-t,c}(u,v)}{|K_{3-t}|} \right| \to \max. \quad (6)$$

Границы первого интервала $[r_u;r_v]^1$ на (5) вычисляются по максимуму критерия (6). Аналогичным образом определим границы для $[r_u;r_v]^q$, $q>1$ на значениях из (5) не вошедших в $[r_u;r_v]^1,\ldots,[r_u;r_v]^{q-1}$. Критерием останова рекурсивной процедуры служит покрытие последовательности (5) множеством из p_c интервалов.

Через $g_{tc}(\mu)$ $(g_{3-t,c}(\mu))$, $t=1,2$ обозначим число значений признака $x_c \in X(n)$ объектов в границах интервала $[r_u; r_v]^\mu$ из класса K_t, (K_{3-t}). Вычислим значение функции принадлежности $f_c(\mu)$ по интервалу $[r_u; r_v]^\mu$ к классу K_1 как

$$f_c(\mu) = \frac{g_{1c}(\mu)/|K_1|}{g_{1c}(\mu)/|K_1| + g_{2c}(\mu)/|K_2|}. \quad (7)$$

По множеству $\{S_j \in K_1 \cup K_2\}_{j=1,\ldots,h}$, $S_j=(x_{j1},\ldots,x_{j11})$ и значениям функции принадлежности (7) определим устойчивость признака $x_c \in X(11)$:

$$U(c) = \frac{1}{h}\sum_{j=1}^{h} \begin{cases} f_c(\mu), x_{jc} \in [r_u;r_v]^\mu, f_c(\mu) > 0,5, \\ 1 - f_c(\mu), x_{jc} \in [r_u;r_v]^\mu, f_c(\mu) < 0,5, \\ 0, x_{jc} \in [r_u;r_v]^\mu, f_c(\mu) = 0,5. \end{cases} \quad (8)$$

Описания набора метаобъектов $R=(Q_1,\ldots,Q_{28})$ по (8) в унифицированном признаковом пространстве формировались по парам классов (K_i,K_j) при $i \neq j$ и $(K_i,E_0 \setminus K_i)$. Каждому $Q \in R$ ставилось в соответствие пара вида (i,j), $i \in \{3,\ldots,9\}$, $j \in \{0,4,\ldots,9\}$. Значение $j=0$ в $(i,0)$ обозначает дополнение $E_0 \setminus K_i$ к классу K_i. Вопрос об

интерпретации результатов кластерного анализа по набору R для практического применения в работе не рассматривается.

Определим локальную область в форме гипершара по радиусу $\varepsilon=\varepsilon(k)$, как $\varepsilon(k)=\frac{\sum_{i=1}^{28}\max_{Q\in O(k,Q_i)}\rho(Q,Q_i)}{28}$, где $O(k,Q_i)$ – гипершар с центром в Q_i, содержащий k ближайших по метрике $\rho(x,y)$ объектов из R. Значение плотности распределения по гипершару с центром в объекте $Q\in R$ вычислим как

$$\varphi(Q,\varepsilon)=\sum_{\rho(Q,Q_i)<\varepsilon}\left(1-\frac{\rho(Q,Q_i)}{\varepsilon}\right). \qquad (9)$$

Присвоение категорий метаобъектам из R при делении их на два класса Z_1 и Z_2 по значению (9) зависело от выбора радиуса ε и границы между классами. При выборе медианы на множестве значений $\{\varphi(Q,\varepsilon)\}$ в качестве границы для проведения вычислительного эксперимента соотношение мощностей классов было $|Z_1|:|Z_2|=14:14$. Подтверждается единственность разбиения метаобъектов на группы по отношению связанности (1). Различие в числе метаобъектов, выбранных в качестве эталонов минимального покрытия в зависимости от используемых мер расстояния демонстрируется в табл. 1. В скобках указано число эталонов покрытия по метрике Евклида.

Таблица 1. Количество групп и эталонов минимального покрытия по локальным метрикам на базе Евклидовой в унифицированном признаковом пространстве

Радиус $\varepsilon(k)$	Класс Z_1		Класс Z_2	
	групп	эталонов	групп	эталонов
ε(2)	3	4(5)	5	4(7)
ε(3)	2	3(5)	5	4(6)
ε(4)	1	1(2)	2	2(2)

Результаты из табл. 1 интерпретируются следующим образом:

– количество непересекающихся групп по классу Z_1 меньше чем в Z_2 с относительно низкой плотностью распределения метаобъектов;

– число эталонов минимального покрытия по локальным метрикам меньше или равно аналогичного числа по метрике Евклида.

Вычисление показателей по (2), (3) продемонстрируем по группе из 12 метаобъектов класса Z_1. Состав класса сформирован по значениям плотности распределения (9) с радиусом ε(3). Влияние выбора эталонов минимального покрытия по локальным метрикам и метрике Евклида на значения $\lambda(Q)$, $\omega(Q)$ показано в табл. 2. В скобках указана идентификация (по оценкам качества вина) объектов классов на E_0, используемых для индексации метаобъектов.

Таблица 2. Значения структурных показателей объектов группы по локальным метрикам и метрике Евклида

№ объекта	Локальные метрики		Метрика Евклида	
	$\lambda(Q)$	$\omega(Q)$	$\lambda(Q)$	$\omega(Q)$
1(4,0)	0.5000	0.7757	0.5000	0.7757
2(5,0)	0.5000	0.1039	0.5000	0.6038
4(7,0)	0.5000	0.2992	0.5000	1.0000
5(8,0)	0.5000	1.0000	0.5000	0.4632
13(4,5)	0.5000	0.2001	0.5000	0.2001
14(4,6)	0.5000	1.0000	0.5000	1.0000
15(4,7)	0.1843	0.3584	0.5000	1.0000
18(5,6)	0.5000	-0.4222	0.5000	0.0116
19(5,7)	0.2594	0.5522	0.5000	1.0000
20(5,8)	0.1976	0.5119	0.5000	0.2226
22(6,7)	0.5000	-.03025	0.5000	0.3930
23(6,8)	0.5000	0.6713	0.5000	0.4128
Среднее по группе	0.4284	0.3957	0.5000	0.5902

Все ближайшие по метрике Евклида граничные объекты (см. табл. 2) из класса Z_2 являются эталонами минимального покрытия так как $|\{\ Q\ |\ \lambda(Q)=0.5\}|=12$. Согласно условия (4) группа является компактной только по метрике Евклида.

V. Заключение

Разработана новая методика оценки качества кластерного анализа с использованием метрических алгоритмов классификации. Для присвоения категорий объектам применялись упорядоченные значения плотности их распределения. В процессе группировки с учётом категорий использовалось отношение связанности объектов по системе пересекающихся гипершаров. Обоснованием выбора этого отношения является:

– единственность числа и состава групп с несферической формой конфигурации;

– возможность получить минимальное покрытие групп эталонами по базовой метрике или локальным метрикам на основе базовой.

Выводы об отделимости групп друг от друга делаются на основе значений оценок, вычисляемых по множеству эталонов покрытия и граничных объектов классов. Предложена технология формирования унифицированного пространства по описаниям наборов объектов из двух классов. Элементами пространства являются значения устойчивости признаков. Одной из важных задач Big Data является кластеризация при наличии пропущенных значений признаков. Классические алгоритмы требуют предварительной импутации или модификаций, тогда как предлагаемая технология унифицированного пространства потенциально позволяет обходиться без импутации: устойчивость признаков можно вычислить даже при неполных данных, и затем выполнять кластеризацию метаобъектов.

Библиография

[1] Игнатьев Н.А., Акбаров Б.Х. Оценка близости структур отношений объектов обучающей выборки на многообразиях наборов латентных признаков // Вестник Томского государственного университета. Управление, вычислительная

техника и информатика. 2023. № 65. С. 69–78. doi: 10.17223/19988605/65/7

[2] Наврузов Э.Р. О формировании баз прецедентов для решения задач информационной безопасности // Вестник РГГУ. «Информатика. Информационная безопасность. Математика» Россия, Москва. 2022. №3. С. 66-84. doi: 10.28995/2686-679X-2022-3-66-84.

[3] Kleinberg J. An Impossibility Theorem for Clustering. https://www.cs.cornell.edu/home/kleinber/nips15.pdf

[4] 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.

[5] Игнатьев Н.А., Згуральская Е.Н. Кластерный анализ с применением обучения на основе отношений связанности и плотности распределения // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2024. № 68. С. 66–74. doi: 10.17223/19988605/68/7

[6] Zhu Y., Ting K.M., Carman M.J. Density-ratio based clustering for discovering clusters with varying densities // Pattern Recognition. 2016. V. 60. P. 983–997.

[7] Айдагулов Р. Р., Главацкий С. Т., Михалёв А. В. Методы осреднения в задачах кластеризации больших данных // Интеллектуальные системы. Теория и приложения. 2021. № 25 (4). С. 12–18.

[8] Bulatov V.G., Alekseev V.P., Vorontsov K.V. Determination of the Number of Topics Intrinsically: Is It Possible? 14 June 2024. https://arxiv.org/pdf/2406.10402.

[9] Сивоголовко Е.В. Методика оценки качества четкой кластеризации // Компьютерные инструменты в образовании. 2011. № 4. С. 14–31.

[10] https://archive.ics.uci.edu/dataset/186/wine+quality, свободный. Яз. англ. (дата обращения 30.11.2025).

Selection of target features for classification and cluster analysis of object relationship structures

N. Ignatev

Abstract. **The article addresses the problem of assessing the quality of clustering objects with an arbitrary (non-spherical) configuration shape. Categories (high, low) are assigned to objects based on their distribution density in local regions of the feature space. A subset of boundary objects relative to objects of another category is determined. The concept of a metaobject is introduced for the case of multimodal data distribution. Feature stability values for objects from two classes corresponding to different modalities are used to describe a metaobject. Stability values across a set of class pairs form a unified space of new features for describing metaobjects. General constraints on the set of their admissible values are explained by the properties of membership functions in fuzzy logic theory. To analyze the clustering quality of metaobjects, they are divided into categories by distribution density. The uniqueness of the number and composition of groups with a non-spherical configuration is ensured by utilizing the connectivity relations of objects of the same category within a system of intersecting hyperspheres. The centers of the hyperspheres are the group objects. At least one boundary object lies at the intersection of two or more hyperspheres. To calculate intra-group proximity and inter-group difference measures, distances to the minimum coverage prototypes within the group and to a subset of boundary objects are used. Clustering quality estimates for each group are derived from the ratio of these measures. A condition on the estimate values for determining group compactness is proposed. A method for assessing the clustering quality of objects with an arbitrary configuration shape is presented, considering distribution density and placement topology. Topological properties are expressed through the connectivity relation of objects depending on their density category. Using real data, their mapping into a unified feature space is demonstrated, followed by an evaluation of the cluster analysis quality for metaobjects. The technology for forming a unified feature space for describing metaobjects can be useful in problems with missing data and in the presence of multimodal distribution.**

Keywords**: clustering quality assessment, minimum coverage standards, feature stability, unified feature space**

REFERENCES

[1] N.A. Ignatev, B.Kh., Akbarov, “Estimation of the proximity of structures of relations of objects of the training sample on manifolds of sets of latent features”, Tomsk State University Journal of Control and Computer Science, no. 65, pp. 69–78. doi: 10.17223/19988605/65/7, 2023. [RUS]

[2] E. R. Navruzov, “On the formation of a precedent base for solving information security problems”, RSUH Bulletin. Computer Science. Information Security. Mathematics., no. 3, pp. 66–84. doi: 10.28995/2686-679X-2022-3-66-84, 2022. [RUS]

[3] J. Kleinberg, An Impossibility Theorem for Clustering. https://www.cs.cornell.edu/home/kleinber/nips15.pdf.

[4] N.A. Ignatyev, “Structure Choice for Relations between Objects in Metric Classification Algorithms”, Pattern Recognition and Image Analysis, vol. 28, no. 4, pp. 590–597, 2018

[5] N.A. Ignatev, E.N. Zguralskaya, “Cluster analysis using learning based on connectivity and distribution density relations”, Tomsk State University Journal of Control and Computer Science, no. 68, pp. 66–74. doi: 10.17223/19988605/68/7, 2024. [RUS]

[6] Y. Zhu, K.M. Ting, M.J. Carman, “Density-ratio based clustering for discovering clusters with varying densities”, Pattern Recognition, vol. 60, pp. 983–997, 2016.

[7] R. R. Aidagulov, S. T. Glavatsky, Mikhalev, “Averaging Methods in Big Data Clustering Problems”, Intelligent Systems: Theory and Applications, no. 25(4), pp. 12–18, 2021. [RUS]

[8] V.G. Bulatov, V.P. Alekseev, K.V. Vorontsov, “Determination of the Number of Topics Intrinsically: Is It Possible?”, https://arxiv.org/pdf/2406.10402, 2024.

[9] E.V. Sivogolovko, “Methodology for assessing the quality of clear clustering”, Computer tools in education, no. 4, pp. 14–31, 2011. [RUS]

[10] https://archive.ics.uci.edu/dataset/186/wine+quality, free. language: English (accessed November 30, 2025).