УДК 519.95

ОТНОШЕНИЕ СВЯЗАННОСТИ В МЕТРИЧЕСКИХ АЛГОРИТМАХ КЛАССИФИКАЦИИ И АНАЛИЗ ЕГО СВОЙСТВ

Игнатьев Н.А., Рамазонов Ш.Ш.

*n_ignatev@rambler.ru

Национальный университет Узбекистана имени Мирзо Улугбека, 100174, Узбекистан, Ташкент, ул. Университетская, 4.

Рассматривается отношение связанности объектов классов для анализа кластерной структуры обучающей выборки. Обсуждается единственность решения задачи о минимальном покрытии обучающей выборки эталонами. Интерес к проблеме единственности связан с использованием в качестве мер расстояния базовой метрики для всех объектов или локальных метрик на основе базовой. Особенность структуры отношений объектов выражается в несферической форме конфигурации кластеров. Для таких кластеров оценок качества не существует. Исследуется свойство отношения связанности и его применение в качестве источника новых знаний при формировании информационных моделей в предметных областях. Свойство связанности предлагается оценивать через поиск наиболее удалённых объектов кластера методом Дейкстра. Исходными данными является матрица смежности, построенная на основе информации о пересечении гипершаров с центрами в объектах кластера. Радиусами гипершаров являлось расстояние до ближайших объектов противоположных классов. Приводятся примеры количественных характеристик кластеров и возможные сферы их применения. Одной из таких характеристик является коэффициент кривизны кластера.

Ключевые слова: отношение связанности, меры компактности, алгоритм Дейкстры, метрические алгоритмы.

Цитирование: *Игнатьев Н.А., Рамазонов Ш.Ш.* Отношение связанности в метрических алгоритмах классификации и анализ его свойств // Проблемы вычислительной и прикладной математики. – 2025. – №4(68). – С. 122-129.

DOI: https://doi.org/10.71310/pcam.4_68.2025.06.

1 Введение

Теоретический основой задач классификации является гипотеза о компактности. Для исследования соответствия реальных задач из предметных областей этой гипотезе предложены специальные меры компактности. Известно, что структура отношений объектов, используемая при реализации метрических алгоритмов распознавания зависит от нескольких факторов. Перечень этих факторов приводится в [4]. Загоруйко Н.Г. [1] была предложена идея о наличии связи между мерами компактности и обобщающей способностью алгоритмов. По его мнению эта связь должна выражаться в наличии некоторого минимального множества объектов, точность распознавания на котором выше чем на исходном множестве.

Практическое воплощение идея Н.Г. Загоруйко нашла в работе [3]. Решение задачи было получено путём формирования базы прецедентов через минимальное покрытие обучающей выборки эталонами. Особенность поиска минимального покрытия и его влияние на обобщающую способность метрических алгоритмов связаны с :

– введением понятия относительного отступа между граничными объектами классов;

– определением необходимых и достаточных условий для отнесения граничного объекта к множеству шумовых;

– вычислением меры компактности по объектам минимального покрытия и определяемыми наборами шумовых объектов;

– введением критериев-регуляризаторов и вычислением обобщающей способности алгоритмов распознавания на его основе.

Повышение точности распознавания на контрольных выборках по базе прецедентов объясняется использованием процедуры вычисления параметров локальных метрик для эталонов минимального покрытия. Показана зависимость этих параметров от изменения состава граничных объектов после удаления шумовых объектов.

Свойства отношений связанности объектов использовались при анализе конфигурации кластеров в [2]. Отношение связанности является средством для изучения топологических структур объектов классов. Необходимо было установить связь единственности разбиения объектов на кластеры с выбором эталонов минимального покрытия множества прецедентов для машинного обучения.

Решение задачи о минимальном покрытии является np-полной. Требовалось определить дополнительные условия для доказательства единственности полученного в [4] решения о покрытии и её эквивалентность задаче с полным перебором всех возможных вариантов.

Дополнительные условия относительно метода динамического программирования заключаются в использовании упорядоченного разбиения задачи на последовательность подзадач. При вычислениях применялось упорядочение по невозрастанию кластеров по их мощности. Методом последовательного исключения объектов кандидаты в эталоны удалялись из состава каждого кластера согласно их расстояниям до граничных объектов из противоположных классов.

Важной характеристикой природы среды обучающей выборки является плотность распределения объектов [5–7]. Известные методы оценки плотности распределения такие как k ближайших соседей и парзеновское окно невозможно применять для групп с несферической конфигурацией. Показано [1], что при анализе структуры отношений объектов можно получить информацию о форме кластера и кратчайшем незамкнутом пути между множеством эталонов покрытия в его составе. Для уточнения составов кластеров впервые было введено понятие шумовой объект. В отличии от задач распознавания шумовой объект рассматривается как кандидат для перехода из состава одного кластера в другой.

Предметом исследования в данной работе является следующее:

– построение множества путей соединяющих любую пару объектов из кластера;

– определение наиболее удалённых объектов кластера с учётом его несферической формы конфигурации;

– определение формы конфигурации кластеров.

2 Постановка задачи и методы её решения

Задана выборка данных $E_0 = \{S_1, \ldots, S_m\}$, разделённая на l ($l \geqslant 2$) непересекающихся подмножества (классы) $K_1, \ldots, K_l$. Каждый объект $S_i \in E_0$ описывается набором разнотипных признаков $X^{(n)} = (x_1, \ldots, x_n)$.

На $E_0 \cup K_q$, $q = 1, \ldots, l$ определена процедура разбиения на кластеры по отношению связанности объектов по метрике $\rho(x, y)$. Объекты $S_i, S_j \in K_q$ считаются связанными между собой ($S_i \leftrightarrow S_j$), если выполняется условие:

$$\{S \in B(E_0, \rho) \mid \rho(S, S_i) < r_i \text{ и } \rho(S, S_j) < r_j\} \neq \varnothing, \tag{1}$$

где r_i (r_j) — расстояние до ближайшего от S_i (S_j) объекта из $\complement K_q$ по метрике $\rho(x, y)$, а

$$B(E_0, \rho) = \left\{ S \in E_0 \;\middle|\; \rho(S_u, S) = \min_{\substack{S_u \in K_q \\ S_v \in \complement K_q}} \rho(S_u, S_v),\ u = 1, \ldots, m \right\}$$

— множество граничных объектов классов, определяемое на E_0.

Множество $G_{qv} \subset K_q$, $c \geqslant 2$, $v \leqslant |K_q|$, представляет группу связанных объектов в классе K_q, если для любых двух объектов существует путь. Объект $S_i \in K_q$ считается несвязанным, если не существует пути $S_i \leftrightarrow S_j$ ни для одного $S_j \neq S_i$, $S_j \in K_q$.

Множество $B(E_0, \rho)$ используется для описания объектов E_0 в пространстве бинарных признаков по таблице $T = \{t_{ij}\}_{m \times d}$, где $d = |B(E_0, \rho)|$. Значение $t_{ij} = 1$, если выполняется условие (1), и $t_{ij} = 0$ в противном случае.

По таблице T определяется минимальное число групп связанных и несвязанных объектов классов. Группы не пересекаются и представлены объектами из K_q или $\complement K_q$.

Требуется:

- Оценить все расстояния между любой парой $(S_i, S_j) \subset G_{qv}$ и построить соединяющие их пути;
- Определить пару $(S_i, S_j) \subset G_{qv}$ с максимальным расстоянием по отношению связанности.

Для решения задачи предлагается использовать алгоритм Дейкстры и метод поиска кратчайшего незамкнутого пути (КНП). Значение КНП является косвенной оценкой плотности распределения в кластере и альтернативой алгоритму DBSCAN[5].

Дополнительные возможности анализа:

- Выявление связи эталонов минимального покрытия при использовании локальных и базовой метрик;
- Определение крайних и максимально удалённых объектов минимального покрытия кластера.

Идея алгоритма Дейкстры — поиск кратчайших путей между объектами $(S_i, S_j) \subset G_v$. Исходными данными для анализа являются строки таблицы T, представляющие объекты кластера, и столбцы, соответствующие граничным объектам.

Анализируется множество путей $\{\Pi(u, v)\}$ между $S_u \leftrightarrow \cdots \leftrightarrow S_v$ и сумма расстояний $\{R(u, v)\}$ между парами $(S_u, S_v) \subset G_q$.

Максимальное расстояние обозначим как:

$$d(S_u, S_v) = \max_{u,v} R(S_u, S_v).$$

Крайними объектами считаем такую пару $(S_i, S_j) \subset G_v$, для которой:

$$d(S_i, S_j) = \max_{\Pi(S_u, S_v)} d(S_u, S_v).$$

Визуальное представление особенностей анализа структуры объектов кластера с несферической конфигурацией показано на рис. 1.

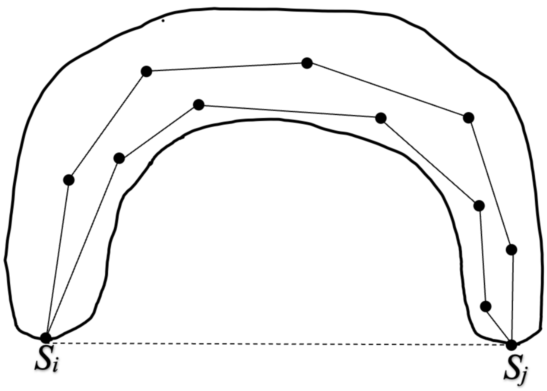

Рис. 1 Особенности вычисления расстояний между объектами кластера по алгоритму Дейкстра

Ценной информацией о кластере является степень её кривизны. Форма конфигурации кластеров может быть различной. Рассмотрим кривизну кластеров в форме подковы.

Пусть (S_i, S_j) — крайние объекты (точки) кластера (см. рис. 1). Определим объект $S_\mu \in G_j$ (см. рис. 2), который является равноудалённым от S_i и S_j, то есть:

$$|d(S_\mu, S_i) - d(S_\mu, S_j)| = \min .$$

Получим треугольник со сторонами $d(S_\mu, S_i)$, $d(S_\mu, S_j)$ и $d(S_i, S_j)$.

Считая, что объект S_μ лежит на сфере, степень крутизны (кривизны) кластера определяется по формуле:

$$k \approx \frac{2\sqrt{1-\left(\frac{d(S_\mu,S_j)^2+d(S_\mu,S_i)^2-d(S_i,S_j)^2}{2\cdot d(S_\mu,S_i)\cdot d(S_\mu,S_j)}\right)^2}}{d(S_i,S_j)}$$

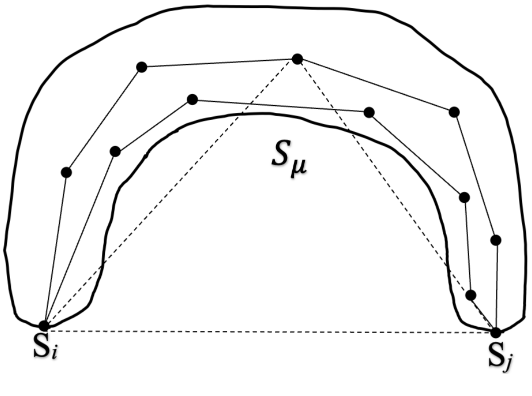

Рис. 2 Анализ крутизны кластера в форме подковы

Отметим многообразие форм конфигурации кластеров по отношению связанности объектов. Предметом исследования может быть изменение формы от выбора метрики, нормирования данных, отбора информативных признаков. Отдельный интерес представляет закономерностей в конкретных предметных областях. Для форм конфигурации, отличных от изображенных на рис 1,2 необходимы специфические методы для анализа.

Технологию анализа кластеров с несферической конфигурацией нельзя автоматически адаптировать для результатов алгоритма DBSCAN. Для алгоритмов NN структура связей рассматривается как топологическая. В случае с DBSCAN изучается свойства плотности распределений объектов в зависимости от их статуса.

3 Вычислительный эксперимент

В качестве данных для эксперимента использовалась выборка German [8]. Выборка представлена 1000 объектами, разделёнными на два класса, мощность которых $|K1| = 700$, $|K2| = 300$. Из 20 признаков в описании объектов 7 измеряются в количественной, 13 в номинальной шкале. Для использования метрики Журавлева в качестве меры расстояния между объектами было произведено отображение значений количественных признаков в [0;1]. Последовательности (цепочки) для связи наиболее удалённых объектов групп по алгоритму Дейкстра представлены в табл.1.

Таблица 1 Результаты анализа 5 групп объектов по алгоритму Дейкстра с учётом их связанности

N / n	**Количество объектов в группах (класс)**	**Последовательность связанных объектов**	**Число путей в графе (к-во рёбер)**	**Мера расстояния**
1	46(1)	0→288→838→438→482→776→893 →851→683→765→27 61→288→838→438→482→776→893 →851→683→765→27 177→288→838→438→482→776→893 →851→683→670→620	39(11)	8.2541 8.2541 9.2644
3	64(1)	737→99→292→939→422→871→110 →990→530→863→294 950→99→292→939→422→871→110 →990→530→863→294	2(11)	4.0110 4.0110
5	54(1)	82→298→537→396→584→843→669 →996→453→338→743→410 82→298→537→396→584→843→669 →996→453→338→743→605 64→279→338→453→996→669 →843→584→396→537→400→82	12(12)	6.0703 6.0703 6.0703

Для исследования отношений связанности объектов в кластере предлагается использовать свойства многообразий структур из теории графов. С этой целью из отношений связанности по таблице T необходимо сформировать матрицу инценденций (смежности) между объектами или вершинами графа. В качестве ожидаемых результатов планируется получить тип графа (например, древовидный) степени вершин, наличие (отсутствие) петель и т.д.

Таблица 2

7	104(1)	146→189→48→705→760→629→8→ 770→365→452→263→262→17→136 → 51→108→130→255→945→549 215→189→48→705→760→629→8→ 770→365→452→263→262→17→136 → 51→108→130→255→945→549	8(20)	26.4859 26.4859
49	19(1)	85→920→269→433→121→977→662 403→920→269→433→121→977→662	4(7)	1.3868 1.3868

Матрица инценденций (смежности) для кластера из 19 объектов Матрица имеет следующий вид, показанный на рис.3.

Свойства матрицы: не симметрична, граф ориентированный (19 вершин, 49 рёбер). Средством для анализа выбран запрос в системе ChatGPT.

Таблица 3 Матрица смежности графа

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	→ 1
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	→ 2
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	→ 1
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	→ 2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	→ 1
0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	→ 2
0	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	→ 9
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	→ 4
0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	→ 3
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	→ 1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	→ 1
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	→ 2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	→ 1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	→ 1
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	→ 3
0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	→ 3
1	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	→ 5
0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	→ 4
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	→ 2

4 Заключение

Предложен методика для оценки кластерной структуры объектов с произвольной (несферической) формой конфигурации. Методика основывается на свойствах отношений связанности объектов по системе пересекающихся гипершаров. Доказана единственность разбиения на кластеры по этим свойством. Определяющим свойством для анализа является наличие путей для связи любых пар объектов кластера.

Из многообразия путей предложено искать кратчайшие. Средством для анализа был использован алгоритм Дейкстры.

Приведены примеры анализа конфигураций кластеров через определение наиболее удаленных объектов. При определении таких объектов учитывалась конфигурация кластеров.

Литература

[1] *Ignatyev N.A.* Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis, – 2028. – Vol. 28. – №4. – P. 590–597.

[2] *Загоруйко Н.Г., Кутненко О.А.* Цензурирование обучающей выборки // Вестник Томского государственного университета. Управление, вычислительная техника и информатика, – 2013. – № 1(22). – С. 66–73.

[3] *Игнатьев Н.А., Турсунмуротов Д.Х.* Цензурирование обучающих выборок с использованием регуляризации отношений связанности объектов классов // Научно-технический вестник информационных технологий, механики и оптики, – 2024. – Т. 24. – №2. – С. 322–329. doi: `http://dx.doi.org/10.17586/2226-1494-2024-24-2-322-329`.

[4] *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: `http://dx.doi.org/10.17223/19988605/68/7`

[5] *Zhu Y, Ting KM.* Carman MJ Density-ratio based clustering for discovering clusters with varying densities // Pattern, – 2016. – V 60. – P. 983–997.

[6] *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.

[7] *Сивоголовко Е.В.* Методы оценки качества чёткой кластеризации // Компьютерные инструменты в образовании. – 2011. – №4. С. 14–31.

[8] *Hastie., Tibshirani R., Friedman J.* The Elements of Statistical Learning:Data Mining,Infrence and Prediction Second Edition (Springer Series in Statistics) . Springer – 2009. – 767 p.

[9] *Рудаков К.В.* О некоторых факторизациях полуметрических конусов и оценках качества эвристических метрик в задачах анализа данных // Доклады Российской Академии наук. Математика, Информатика, Процессы Управления. – 2020. – С. 101–103.

[10] *Зухба А.В.* Оценка вычислительной сложности задач отбора эталонных объектов и признаков // диссертация кандидата наук [Электронный ресурс]. Режим доступа: https://istina.ips.ac.ru/dissertations/ (дата обращения: 28.02.2025).

[11] https://archive.ics.uci.edu/dataset/94/spambase, свободный. Яз. англ. (дата обращения: 23.06.2025).

[12] *Кобринский Б.А.* Доверие к технологиям искусственного интеллекта // Искусственный интеллект и принятие решений. – 2024. – №3. – С. 3–17.

[13] *Боженюк А.В., Беляков С.Л., Косенко О.В.* Оценка информационной надежности сложных систем с помощью интуиционистских нечетких графов // Наука и технология железных дорог. – 2019. – Т.3. – №4(12). – С. 65–74.

UDC 519.95

RELATIONSHIP IN METRIC CLASSIFICATION ALGORITHMS AND ANALYSIS OF ITS PROPERTIES

Ignatiev N.A., Ramazonov Sh.Sh.
*n_ignatev@rambler.ru
National University of Uzbekistan named after Mirzo Ulugbek,
4, University str., Tashkent 100174, Uzbekistan

The relationship between class objects is considered for analyzing the cluster structure of the training dataset. The uniqueness of the solution to the problem of minimal coverage of the training set by prototypes is discussed. Interest in the uniqueness issue is associated with the use of either a base metric as a distance measure for all objects or local metrics derived from the base one. The specific structure of object relationships is reflected in the non-spherical shape of the cluster configurations. For such clusters, no established quality assessments exist. The property of connectivity is explored and its application as a source of new knowledge during the construction of information models in specific subject areas is studied. The connectivity property is proposed to be evaluated through the search for the most distant cluster objects using Dijkstra's algorithm. The input data is an adjacency matrix constructed based on the intersection information of hyperspheres centered at the cluster objects. The radii of these hyperspheres are defined as the distances to the nearest objects of the opposite classes. Examples of quantitative characteristics of clusters and their possible areas of application are provided. One such characteristic is the cluster curvature coefficient.

Keywords: connectivity relation, compactness measures, Dijkstra algorithm, metric algorithms.

Citation: Ignatiev N.A., Ramazonov Sh.Sh. 2025. Relationship in metric classification algorithms and analysis of its properties. *Problems of Computational and Applied Mathematics.* 4(68): 122-129.

DOI: https://doi.org/10.71310/pcam.4_68.2025.06.

№ 4(68) 2025

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

RAQAMLI TEXNOLOGIYALAR VA
SUN'IY INTELLEKTNI RIVOJLANTIRISH
ILMIY-TADQIQOT INSTITUTI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 4(68) 2025

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Регистрационное свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(998) 712-319-253, 712-319-249.
Э-почта: `journals@airi.uz`.
Веб-сайт: `https://journals.airi.uz`.

Дизайн и вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 29.08.2025 г.
Формат 60x84 1/8. Заказ №6. Тираж 100 экз.

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

No. 4(68) 2025

The journal was established in 2015.
6 issues are published per year.

Founder:
Digital Technologies and Artificial Intelligence Development Research Institute.

The journal is registered by Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan.
The registration certificate No. 0856 of 5 August 2015.

ISSN 2181-8460, eISSN 2181-046X

Address:
100125, Tashkent, Buz-2, 17A.
Tel.: +(998) 712-319-253, 712-319-249.
E-mail: journals@airi.uz.
Web-site: https://journals.airi.uz.

Layout design:
Sharipov Kh.D.

DTAIDRI printing office.
Signed for print 29.08.2025
Format 60x84 1/8. Order No. 6. Print run of 100 copies.

Содержание

Contents