УДК 519.254

Упрощенный показатель силуэта для определения качества кластерных структур

В.В. Журавлева, А.С. Маничева

Алтайский государственный университет (Барнаул, Россия)

Simplified Silhouette Parameter for Assessing the Quality of Cluster Structures

V.V. Zhuravleva, A.S. Manicheva

Altai State University (Barnaul, Russia)

Обсуждаются вопросы, связанные с оценкой качества построения кластерной структуры данных. Приведено описание показателя качества кластеризации, учитывающего характеристики компактности и отделимости кластеров, — показателя силуэта в двух вариантах: классического и упрощенного. Отмечено, что для вычисления классического показателя силуэта на большом массиве данных требуется трудоемкая процедура полного перебора пар объектов. Предложена вариация данного показателя, удобная для оценки кластерных структур, построенных на больших массивах данных, — упрощенный показатель силуэта. Рассмотренный показатель протестирован на модельных данных, по которым было построено несколько вариантов кластерных структур, таких, что отдельные кластеры представляли совокупность мини-кластеров. В качестве объектов при вычислении внутрикластерных и межкластерных расстояний были выбраны центры мини-кластеров с учетом их «веса» (в качестве веса задавалось число объектов в мини-кластерах). По каждой кластерной структуре тестового набора данных был вычислен соответствующий показатель силуэта. Проведенное сравнение значений классического и упрощенного показателей силуэта для каждого набора модельных данных дало адекватную оценку качества кластеризации.

Ключевые слова: кластерный анализ, кластерная структура, качество кластеризации, показатель силуэта, компактность, отделимость.

The article deals with issues related to assessing the quality of a cluster data structure. A description of the clustering quality index is given, which takes into account the characteristics of compactness and separability of clusters in two versions: the classical and the simplified silhouette index. It is noted that a laborious procedure of a complete enumeration of pairs of objects is required to evaluate the classical silhouette feature on big data. Further, a variation of this indicator — a simplified silhouette indicator — is proposed and found to be convenient for assessing cluster structures built on big data arrays. The sample indicator has been tested on model data, and several variants of cluster structures are built for the objects like identified clusters that are present in the set of mini-clusters. The centers of mini-clusters with consideration to their "weight" (the number of objects in mini-clusters was set as the weight) are chosen as objects when calculating intra-cluster and inter-cluster distances. The corresponding silhouette parameter is calculated. The comparison of the indicators of the classical and simplified silhouette indicators for each set of data models provides an adequate assessment of the quality of clustering.

Key words: cluster analysis, cluster structure, clustering quality, silhouette parameter, compactness, separability.

DOI: 10.14258/izvasu(2022)4-17

1. Вводные замечания

Методы кластеризации предназначены для разбиения совокупности объектов на однородные группы (кластеры или классы). Кластер можно охарактеризовать как группу объектов, имеющих общие свойства. К основным характеристикам «хороших» кластерных структур принято относить:

- внутреннюю однородность объектов в кластере, или компактность (объекты одного кластера должны иметь большое сходство);
- изолированность объектов разных кластеров, или отделимость (объекты разных кластеров должны иметь малое сходство) и компактность.

Если данные представить как точки в признаковом пространстве, то задача кластеризации сводится к определению «сгущений точек». Процедура кластеризации является описательной, она не делает статистических выводов, но дает возможность провести разведочный анализ и изучить «структуру данных» [1–3].

Кластерный анализ применяется в различных областях [4–8]. Технологии кластеризации исключительно полезны, когда нужно проанализировать структуру массивов больших данных непосредственно либо как вспомогательные методы при снижении размерности данных.

Следует отметить, что в результате применения различных методов кластерного анализа могут быть получены кластеры различной формы. Например, возможны кластеры удлиненной формы либо ситуация, когда кластеры представлены длинными «цепочками» и т.д., а некоторые методы могут создавать кластеры произвольной формы. Кроме того, результат кластеризации одним и тем же методом может изменяться в зависимости от выбранной метрики расстояния либо процедуры стандартизации данных [1–4].

Оценка качества кластерной структуры может проводиться на основе следующих процедур [1–3]:

• экспертная проверка;

• выбор контрольных объектов и проверка их на полученных кластерах;

• определение стабильности кластерной структуры после добавления новых признаков;

• сравнение кластерных структур, построенных с использованием различных методов либо одного метода с разными параметрами.

В последнем случае получение различными методами схожих кластеров указывает на «успешность» кластеризации. Иначе для выбора лучшего результата используются, как правило, относительные показатели качества кластеризации, которые оценивают качество, сравнивая несколько кластерных структур между собой, не имея априорной информации и принимая в расчет только сведения о кластерной структуре. Среди прочих используется и коэффициент силуэта. Подробный обзор различных показателей качества кластеризации приведен в работе [9].

Для вычисления классического показателя силуэта на большом массиве данных требуется трудоемкая процедура полного перебора пар объектов. Целью работы является разработка варианта данного показателя, удобная для оценки кластерных структур, построенных на больших массивах данных, — упрощенный показатель силуэта.

2. Классический показатель силуэта

Силуэт каждого кластера определяется следующим образом [9–10]: пусть объект x_j принадлежит кластеру c_p. Обозначим среднее расстояние от этого объекта до других объектов из того же кластера c_p через a_{p_j}. Теперь обозначим среднее расстояние от x_j до объектов из другого кластера c_q ($q \neq p$) через d_{q_j}. Зададим

$$b_{p_j} = \min d_{q_j} \tag{1}$$

как меру несхожести выбранного объекта с ближайшим кластером. Таким образом, силуэт каждого отдельного объекта определяется по формуле

$$S_{x_j} = \frac{b_{p_j} - a_{p_j}}{\max\left(a_{p_j}, b_{p_j}\right)}. \tag{2}$$

Значения показателя силуэта ограничены отрезком [–1; 1]. Очевидно, что высокое значение показателя S_{x_j} характеризует собой «лучшую» принадлежность объекта x_j к кластеру p.

Силуэтом кластера называется средняя величина показателя силуэта всех объектов кластера [9–10]. Таким образом, силуэт показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров. Данная величина также лежит в диапазоне [–1,1]. Значения, близкие к –1, соответствуют неудачным кластерным структурам; значения, близкие к нулю, говорят о том, что кластеры пересекаются или накладываются друг на друга; значения, близкие к 1, соответствуют «плотным» четко выделенным кластерам. Таким образом, чем больше силуэт, тем более четко выделены кластеры, и они представляют собой компактные, плотно сгруппированные облака точек.

Оценка для всей кластерной структуры достигается усреднением показателя по всем объектам N [9–10]:

$$SWC = \frac{1}{N}\sum\nolimits_{j=1}^{N} S_{x_j} . \tag{3}$$

Лучшее разбиение характеризуется наибольшим значением показателя *SWC*, что достигается в том случае, когда расстояния внутри кластеров a_{p_j} малы, а расстояния между элементами соседних кластеров b_{p_j} велики.

Показатель силуэта зависит от формы кластеров и достигает больших значений на более выпуклых кластерах, получаемых с помощью алгоритмов, основанных на восстановлении плотности распределения (в том числе при применении алгоритма К-средних к «хорошим» данным). Если же структура состоит из близкорасположенных кластеров очень сложной формы (например, при использовании алгоритма DBSCAN), то значение показателя силуэта невелико и он не всегда отражает реальную ситуацию.

3. Упрощенный показатель силуэта для больших массивов данных

Определение средних внутрикластерных расстояний и меры несхожести для каждого объекта с ближайшим кластером по формуле (1) на большом массиве данных требует проведения полного перебора пар объектов, что может занять время на несколько порядков большее, чем требуется для выполнения самого алгоритма кластеризации.

Проблема времени может быть решена в данном случае двумя способами:

• многократное распараллеливание вычислений при подсчете средних расстояний до разных кластеров (так следует поступить для каждого объекта);

• разумное упрощение показателя силуэта, при котором не требуется полный перебор объектов.

В некоторых работах в качестве меры несхожести предлагается брать расстояния от каждого объекта до центров других кластеров и вычислять упрощенный индекс силуэта по прежнему алгоритму [10]. Такой подход все же требует полного перебора и дает адекватный результат лишь для структуры с выпуклыми хорошо разделенными кластерами, а во многих других ситуациях может давать значительное искажение показателя.

Здесь предлагается для оценки кластерных структур больших массивов данных применять следующий подход: при оценке силуэта кластеров учитывать не все объекты, а только эталонные представители (для большей точности таковых эталонов должно быть достаточно много). Это возможно в следующих ситуациях. Случай 1 — кластерная структура построена в два этапа: на первом этапе построено разбиение на большое количество мини-кластеров малого размера (с очень высокой степенью сходства объектов) при помощи любого подходящего алгоритма; на втором этапе произведено их объединение в большие кластеры. Случай 2 — кластерная структура построена непосредственно алгоритмом, позволяющим получать кластеры сложной формы, поэтому для упрощения их описания каждый кластер дополнительно разбивается на мелкие «скопления», центры которых объявляются эталонами.

Алгоритм, соответствующий случаю 1, описан в работах [11–12]. К случаю 2 можно отнести вариации популярного алгоритма Форел [2].

Итак, пусть имеется N сложных кластеров C_k, каждый из которых состоит из n_k мини-кластеров $c_{k1},\ldots,c_{kn_k} \in C_k$. Пусть мини-кластер c_{kn_j} описан центром z_{kn_j} и количеством входящих объектов m_{kn_j} (рис. 1).

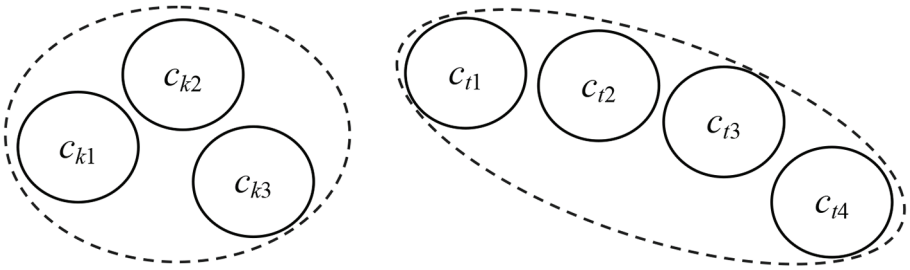

Рис. 1. Пример сложных кластеров

Будем определять силуэты мини-кластеров s_{kn_j} по формуле, аналогичной (2), где в качестве объектов при вычислении расстояний берутся центры z_{kn_j} мини-кластеров с учетом их «веса» m_{kn_j}. Однако в процессе вычислений возникают отличия при определении величин a_{p_j} и d_{q_j}. Последнюю в данном случае надо понимать буквально, как взвешенное среднее расстояние от центра выбранного мини-кластера до центров другого кластера. Величина a_{p_j} будет средневзвешенным расстоянием до центров мини-кластеров внутри того же кластера. Если же кластер состоит из одного мини-кластера, то внутрикластерное расстояние в нем очень мало, а при выбранном упрощенном вычислении будет в точности равно нулю.

Тогда формула для силуэта кластера C_k примет вид:

$$S_k = \frac{1}{N_k}\sum\nolimits_{j=1}^{k} s_{kn_j} m_{kn_j} . \qquad (4)$$

где $N_k = \sum\nolimits_{j=1}^{k} m_{kn_j}$ — количество объектов в выбранном кластере.

Силуэт кластерной структуры будет получен как усреднение силуэтов кластеров

$$S = \frac{1}{N}\sum\nolimits_{k=1}^{n_k} S_k N_k . \qquad (5)$$

В качестве замечаний к вышеизложенному следует указать:

• при вычислении показателя силуэта в качестве функции расстояния целесообразно брать ту же метрику, которая была использована при построении кластерной структуры;

• если некоторый кластер состоит из одного мини-кластера, то значение показателя силуэта для указанного кластера будет равно 1;

• описанный подход в некоторой степени учитывает «плотность» точек внутри мини-кластеров, что вполне «справедливо» при упрощении оригинальной формулы показателя силуэта (т.е. усредненные расстояния как бы вычисляются для всех объектов исходного массива данных).

4. Тестирование

Описанный выше упрощенный показатель силуэта был протестирован на десяти наборах модельных данных. Базовый набор данных был сгенерирован в виде скопления 22 непересекающихся мини-кластеров, в каждом из которых от 1000 до 2000 объектов (рис. 2). Общее количество объектов было задано 30000.

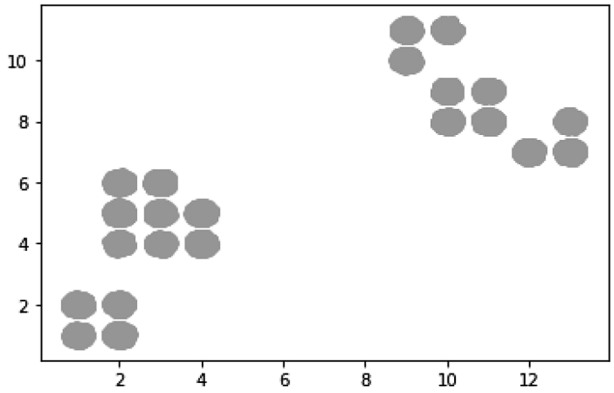

Рис. 2. Базовый набор данных

С использованием алгоритма К-средних было построено разбиение каждого набора модельных данных на большое число мелких кластеров (от 20 до 80). Затем путем объединения различных комбинаций мини-кластеров было получено десять кластерных структур (вариантов тестирования алгоритма расчета упрощенного показателя силуэта) с разным числом кластеров. Результаты расчета представлены в таблице.

Расчетные значения показателя силуэта

Вариант тестирования	Показатель силуэта	
	классическая формула	упрощенная формула
1	0,195	-0,540
2	0,196	-0,494
3	0,198	-0,452
4	0,220	0,089
5	0,245	0,070
6	0,277	0,124

Вариант 1 соответствует наихудшей кластерной структуре (рис. 3), вариант 10 — наилучшей (рис. 4).

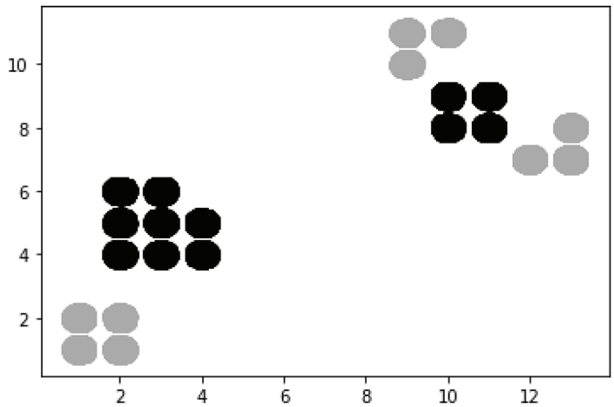

Рис. 3. Наихудшая кластерная структура

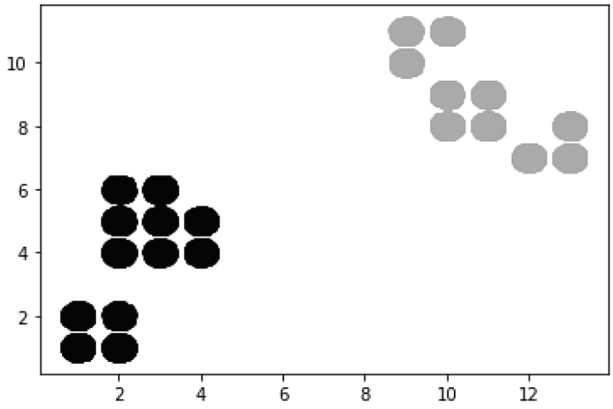

Рис. 4. Наилучшая кластерная структура

Варианты тестирования упорядочены по значению показателя силуэта по классической формуле (3) (по возрастанию). Стоит отметить, что при переходе к упрощенной формуле (5) сохраняется упорядоченность значений показателя силуэта.

Применение упрощенной формулы приводит к значительному снижению временных затрат. Для использованных модельных данных экономия времени составила от 68 % до 81 %.

Заключение

Описанный вариант вычисления упрощенного показателя силуэта позволяет быстро и эффективно оценить качество кластеризации.

Существенным моментом при сравнении различных вариантов разбиений на кластеры (для некоторого выбранного набора данных) является относительная упорядоченность значений показателя силуэта кластерной структуры, которая не изменяется при использовании упрощенного подхода.

Очевидно, что использование упрощенного показателя силуэта целесообразно при анализе кластерных структур больших массивов данных. Для достижения большей экономии времени возможно применение технологий распараллеливания вычислений.

Использование упрощенного показателя силуэта для данных с шумом и для данных с бинарными признаками (как и оригинального индекса силуэта) требует дополнительного исследования.

Библиографический список

1. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск, 1999.

2. Загоруйко Н.Г. Интеллектуальный анализ данных, основанный на функции конкурентного сходства // Автометрия. 2008. Т. 44. № 3.

3. Миркин Б.Г. Методы кластер-анализа для поддержки принятия решений: обзор. М., 2011.

4. Dronov S.V., Evdokimov E.A. Post-hoc cluster analysis of connection between forming characteristics // Model Assisted Statistics and Applications. 2018. Vol. 13. № 2. DOI: 10.3233/MAS-180429.

5. Журавлева В.В., Аюпов К.Е. Применение метода кластерного анализа для обнаружения зависимости обострений сердечно-сосудистых заболеваний от геофизических факторов : сб. научн. ст. Междунар. конф. «Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования». Барнаул, 2015.

6. Айдинян А.Р., Цветкова О.Л. Алгоритмы кластерного анализа для решения задач с асимметричной мерой близости // Сиб. журн. вычисл. матем. 2018. Т. 21. № 2. DOI: 10.15372/SJNM20180201.

7. Игнатьев Н.А. Кластерный анализ данных и выбор объектов-эталонов в задачах распознавания с учителем // Вычислительные технологии. 2015. Т. 20. № 6.

8. Савченко Т.Н. Применение методов кластерного анализа для анализа данных психологических исследований // Прикладная юридическая психология. 2008. № 4.

9. Сивоголовко Е.В. Оценка качества кластеризации в задачах интеллектуального анализа данных : дис. … канд. физ.-мат. наук. СПб., 2014.

10. Паклин Н.Б., Орешков В.И. Кластерные силуэты // Системный анализ в проектировании и управлении : сб. научн. тр. XX Междунар. науч.-практич. конф. Ч. 2. СПб., 2016.

11. Журавлева В.В., Бондарева А.А. Описание одного алгоритма кластеризации типа Forel // МАК-2015 : сб. трудов 18-й Всеросс. конф. по математике. Барнаул, 2015.

12. Журавлева В.В. Об одном алгоритме кластеризации : сб. научн. ст. Междунар. конф. «Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования». Барнаул, 2015.