Алгоритмы устойчивой кластеризации на основе индексных функций и функций устойчивости[1]

Д. С. Шалымов
Санкт-Петербургский государственный университет

Кластеризация активно изучается в таких областях, как статистика, распознавание образов, машинное обучение и др. В работе дается определение понятий кластеризации и устойчивости кластеризации, описывается актуальность и основные проблемы кластеризации, предлагается обзор существующих алгоритмов устойчивой кластеризации на основе индексов и на основе функций устойчивости.

Ключевые слова: кластеризация, устойчивость кластеризации.

1. Введение

Кластеризацией является разбиение множества данных на группы по схожим признакам. Кластеризация используется при решении многообразных задач обработки данных, в том числе при распознавании образов, машинном обучении, автоматической классификации, выработке стратегий управления и т. д.

До сих пор не было найдено какого-либо универсального алгоритма, который был бы эффективным на данных различной природы. В основном используются итеративные методы кластеризации, которые базируются на априорном задании количества кластеров и некотором выборе первоначального разбиения. При этом результат их применения существенно зависит от правильности оценки количества кластеров.

Устойчивость кластеризации показывает, насколько различными получаются результирующие разбиения на группы после многократного применения алгоритмов кластеризации для одних и тех же данных. В данной статье приводится краткий обзор основных

методов, позволяющих оценить устойчивость кластеризации, которая связана с действительным количеством кластеров. Описаны методы на основе индексов, которые сравнивают внутренние и внешние дисперсии кластеров. Также описаны алгоритмы, использующие функции устойчивости, которые определяют соответствие назначенных кластеров для выборочных элементов множества.

Вычислительная сложность известных алгоритмов исследования устойчивости кластеризации существенно растет при увеличении мощности исследуемого множества данных. Также большинство из них недостаточно математически обоснованы. В статье рассматривается несколько помехоустойчивых алгоритмов, которые могут работать на множествах произвольной структуры.

2. Задача кластеризации

Кластеризацию можно определить как процесс объединения данных в группы по схожим признакам. Эта задача является одной из фундаментальных в области анализа данных и Data Mining. Список областей, в которых применяется кластеризация, очень широк: сегментация изображений, прогнозирование, анализ текстов, сжатие данных и многие др. На современном этапе кластеризация часто выступает первым шагом при анализе данных. После выделения схожих групп применяются другие методы. Для каждой группы строится отдельная модель. Решения задач кластеризации используются в таких научных направлениях, как статистика, распознавание образов, оптимизация, машинное обучение, финансовая математика, автоматическая классификация, выработка стратегий управления, исследование свойств ДНК, моделирование филогении организмов (кладистический анализ) и др. Отсюда многообразие синонимов понятию кластер — класс, таксон, сгущение. Однако стоит различать классификацию и кластеризацию. Классификацией называется отнесение каждого элемента в определенный класс с заранее известными параметрами, полученными на этапе обучения. При этом число классов строго ограничено. Кластеризация — это разбиение множества данных на кластеры. Кластерами будем называть подмножества, параметры которых заранее неизвестны. Количество кластеров может быть произвольным или фиксированным.

Цели кластеризации могут быть различными в зависимости от особенностей конкретной прикладной задачи:

— определить структуру множества данных, разбив его на группы схожих объектов, упростив дальнейшую обработку данных в каждом кластере в отдельности;

— сократить объем хранимых данных, оставив по одному наиболее типичному представителю от каждого кластера;

— выделить нетипичные объекты, которые не подходят ни к одному из кластеров.

Основная суть алгоритмов кластеризации заключается в следующем. Имеется обучающая последовательность (набор данных) $\{x_1, \ldots, x_n\} \in X$ и функция расстояния между объектами $\rho(x, x\prime)$. Требуется разбить последовательность на непересекающиеся подмножества (называемые кластерами) так, чтобы каждый кластер состоял из объектов, близких по метрике ρ , а объекты разных кластеров существенно отличались. Алгоритм кластеризации — это функция $a : X \to Y$, которая любому объекту $x \in X$ ставит в соответствие метку кластера $y_i \in Y$. Множество меток Y заранее неизвестно.

На сегодняшний момент число методов разбиения групп объектов на кластеры довольно велико - несколько десятков алгоритмов и еще больше их модификаций.

В кластеризации выделяют два основных подхода: декомпозиция (неиерархический), когда каждый объект связан только с одной группой, и кластеризация на основе иерархий (иерархический), когда каждая группа большего размера состоит из групп меньшего размера.

Классические иерархические алгоритмы работают только с категорийными атрибутами, когда строится полное дерево вложенных кластеров. Здесь распространены агломеративные (объединительные) методы построения иерархий кластеров. В них производится последовательное объединение исходных объектов и соответствующее последовательное уменьшение числа кластеров. Также существуют разделительные техники, когда кластеры разделяются. При этом изначально предполагается, что в системе только

один кластер. Иерархические алгоритмы обеспечивают сравнительно высокое качество кластеризации. Большинство из них имеют сложность $\mathcal{O}(n^2)$.

Неиерархические алгоритмы основаны на оптимизации некоторой целевой функции, определяющей оптимальное в определенном смысле разбиение множества объектов на кластеры. В этой группе популярны алгоритмы семейства k-средних (k-means), которые в качестве целевой функции используют сумму квадратов взвешенных отклонений координат объектов от центров искомых кластеров. Кластеры ищутся сферической либо эллипсоидной формы.

Решение задачи кластеризации принципиально неоднозначно, поскольку не существует однозначно наилучшего критерия качества кластеризации, число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием, а также результат кластеризации во многих алгоритмах существенно зависит от метрики, выбор которой чаще всего субъективен и определяется экспертом.

Таким образом, не существует единого универсального алгоритма кластеризации. При использовании любого алгоритма важно понимать его достоинства и недостатки, учитывать природу данных, с которыми он лучше работает и способность к масштабируемости.

3. Устойчивая кластеризация. Индексные методы

Устойчивость кластеризации является характеристикой, показывающей, насколько различными получаются результирующие разбиения на группы после многократного применения алгоритмов кластеризации для одних и тех же данных. Небольшое расхождение результатов интерпретируется как высокая устойчивость. Количество кластеров, которое максимизирует кластерную устойчивость, может служить хорошей оценкой для реального количества кластеров. Проблема определения устойчивой кластеризации является одной из наиболее сложных проблем кластерного анализа.

Большая часть работ, посвященных устойчивой кластеризации, нацелены на конкретные прикладные проблемы. Известные алго-

ритмы устойчивой кластеризации требуют определенных предположений о своих параметрах и недостаточно математически обоснованы. Также их вычислительная сложность существенно растет при увеличении мощности исследуемого множества данных.

Еще одной существенной трудностью кластерного анализа является нахождение первоначальных центров предполагаемых кластеров, т. е. данных, с которых начинают работу алгоритмы. С теоретической точки зрения это могут быть любые точки пространства. Однако на практике выбор начальных центров значительно сказывается как на скорости сходимости алгоритмов, так и на качестве результатов.

Существует несколько базовых подходов к определению количества кластеров в множестве данных. Они основаны на:

1) определяемых с помощью индексов, сравнивающих степени "разброса" данных внутри кластеров и между кластерами;

2) расчете значений характеристик (функций устойчивости), показывающих соответствие назначенных кластеров для выборочных элементов множества;

3) статистиках, определяющих наиболее вероятное решение;

4) оценивании плотностей распределений.

Рассмотрим наиболее известные алгоритмы устойчивой кластеризации на основе индексов.

Первый подход (Calinski-Harabasz) [1] выбирает количество кластеров как значение аргумента, максимизирующего функцию $CH(K)$,

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)},$$

где $B(K)$ и $W(K)$, соответственно, внешняя и внутренняя суммы квадратов элементов данных с K кластерами. Это один из самых первых предложенных методов. Он оказывается эффективным при данных небольших размерностей.

Подход Krzanowski and Lai [2] максимизирует функцию $KL(K)$:

$$KL(K) = \left|\frac{DIFF(K)}{DIFF(K+1)}\right|,$$

где $DIFF(K) = (K-1)^{2/p}W(K-1)-(K)^{2/p}W(K)$, p — размерность пространства.

Основная идея заключается в измерении порядка изменчивости внутренних дисперсий.

Hartigan [3] предлагает выбирать такое наименьшее значение K, что $H(K)$ меньше либо равно 10

$$H(K) = (n - K - 1)[\frac{W(K)}{W(K+1)} - 1].$$

Еще один метод был предложен Kaufman and Rousseeuw [4], в котором измеряется, насколько i-я точка была хорошо кластеризована. Для этого определяют функцию

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

где $a(i)$ — среднее расстояние между i-ой точкой и всеми остальными наблюдениями, попавшими в тот же кластер, $b(i)$ — среднее расстояние до точек в ближайшем кластере, где под ближайшим кластером понимается тот, который минимизирует $b(i)$. Количество кластеров считается верным, если оно максимизирует среднее значение $s(i)$.

Другой, более редкий, метод использует статистику расхождений [5], вычисляя

$$GAP(K) = \frac{1}{B}\sum_{b} \log W_b^*(K) - \log W(K).$$

Метод использует B различных унифицированных множеств, каждое из которых состоит из того же количества элементов, что и оригинальное множество. Строится внутренняя сумма квадратов для различного количества кластеров. Требуется подобрать K, максимизирующее $GAP(K)$.

Существуют методы, основанные на использовании ядер для вероятностных распределений элементов внутри кластеров [6], которые оперируют с двумя выборками элементов. Выборки производятся согласно двум распределениям. Первое — это исходное распределение данных, второе построено таким образом, что оно представляет ядра кластеров. Качество кластера считается по мере схожести его со своим ядром.

Известен эффективный непараметрический метод, предложенный Sugar and James [7]. Он основан на использовании "искажений", которые по своей сути являются оценками дисперсии внутри класса. Для иллюстрации эффективности алгоритма строится специальный функционал зависимости минимального "искажения" от количества кластеров, который убывает с ростом количества кластеров. Минимальное "искажение" определяется следующим образом: на множестве данных с учетом текущего количества кластеров запускается какой-либо алгоритм кластеризации. Как правило, это k-means. Множество разбивается на отдельные кластеры, для каждого из которых подсчитывается внутренняя дисперсия. Из множества значений дисперсий выбирается наименьшее, которое принимается за минимальное "искажение" и становится значением функционала при данном количестве кластеров. Теоретически и эмпирически доказывается, что, при определенном выборе параметра Y (степень трансформации), построенная для функционала кривая будет иметь резкий скачок в том месте, которое соответствует действительному количеству кластеров.

Процедура определения количества кластеров состоит из следующих шагов:

1. Запускается k-means алгоритм для K кластеров и определяется соответствующее "искажение" $\widehat{d}_k$. Для различных значений K строится набор $\widehat{d}_k$.

2. Выбирается степень трансформации $Y > 0$ (обычно принимается $Y = p/2$).

3. Вычисляются скачки по формуле $J_k = \widehat{d}_k^{-Y} - \widehat{d}_{k-1}^{-Y}$.

4. За итоговое количество кластеров выбирается то, которое соответствует наибольшему скачку $K^* = \arg\max_k J_k$.

На рис. 1 изображены зависимости "искажений" от количества кластеров. В данном случае использовано Гауссово распределение, однако метод был апробирован также на негауссовых данных.

4. Подходы на основе функций устойчивости

Функции устойчивости, как правило, основаны на подсчете количества различий в метках, задаваемых элементам множества на

Рис. 1: a) Набор из девяти кластеров (Гауссово распределение) b) Кривая функционала (внутренняя дисперсия - количество кластеров) c) Преобразованная кривая, показывыающая количество кластеров, равное девяти. d) Кривая скачков.

каждой итерации алгоритма. Рассмотрим некоторые из них.

Основная идея метода, предложенного в [8], заключается в использовании вычислительной процедуры разбиения на группы, которая работает эффективно, когда она идентифицирует схожие структуры за счет многократного применения алгоритмов кластеризации. Выборки производятся за счет произвольного выбора элементов исходного множества. Первая выборка, Y, считается обучающим множеством. На основе этой выборки анализируется тестовое подмножество. Пусть $\alpha_{k,Y}(z)$ и $\alpha_k(z)$ определяют метки элементов $z \in Z$ на основе анализа выборки Y и на основе непосредственного применения алгоритма кластеризации соответственно.

Например, $\alpha_{k,Y}(z)$ может быть реализован с помощью определения каждого элемента из Z к ближайшему центру кластера в Y. Разница между метками используется функцией устойчивости для вычисления структуры кластеров. Так как не используется заранее известных меток для кластеров, значения меток могут быть перемешаны от одной реализации к другой. Таким образом, вводится функция перестановки ψ. Эта функция соотносит индексы между метками так, что соотношение между метками максимально. Для решения этой задачи используется Венгерский метод [9], сложность которого $\mathcal{O}(k^3)$. В данном случае устойчивость определяется сле-

дующим образом:

$$S(\alpha_k) = \max_{\psi} \sum_{z \in Z} I(\alpha_{k,Y} \neq \psi(\alpha_k(z))).$$

Функция устойчивости зависит от k. За результирующее количество кластеров принимается

$$k_* = \min_k S(\alpha_k).$$

Алгоритм в общем виде можно описать следующим образом.

1) Из исходного множество производятся две независимые выборки, в результате чего формируются множества O_1^n и O_2^n, равномощные исходному.

2) Применяется алгоритм α_1 к первому множеству O_1^n, результатом чего является разбиение множества на k кластеров.

3) Применяется алгоритм α_2 ко второму множеству O_2^n. Алгоритм α_2 используется для предсказания кластерной принадлежности всех элементов первого множества O_1^n

4) Каждый элемент множества O_1^n теперь имеет две маркировки. С помощью Венгерского метода [9] определяется корректная маркировка всех элементов множества O_1^n. В качестве стоимости идентификации меток i и j принимается количество несовпадений по сравнению с маркировкой, произведенной в ходе алгоритма $\alpha_1(O_1^n)$, которая считается корректной.

5) Подсчитывается среднее значение “нестабильности” при заданном количестве кластеров k на основе стоимости идентификации меток.

6) За итоговое количество кластеров принимается такое k, которое минимизирует “нестабильность”.

Алгоритмы α_1, α_2 и множества O_1^n, O_2^n могут быть различны. Но, как правило, используется один алгоритм кластеризации α и множества O_1^n, O_2^n содержат все элементы исходного множества и равномощны ему.

Используя два независимых множества из одного и того же источника данных, задача кластеризации переформулируется как задача обучения с учителем, где в качестве учителя выступает разбиение на кластеры первого множества.

Следующий алгоритм [10] применяется для данных, которые содержат в себе помехи. Может применяться для сравнения различных алгоритмов. Помехи обрезаются за счет передискретизации исходного множества данных.

Используются матрицы кластерной связности

$$T_{i,j} = \begin{cases} 1 & i,\ j \in SameCluster, \\ 0 & otherwise. \end{cases}$$

Функция устойчивости основана на сравнении матриц связности всех выбранных подмножеств и исходного множества. Производится следующее: сперва выделяются для одного разбиения все такие элементы, которые попали в один кластер как в данном разбиении, так и в исходном множестве. Далее производится выделение таких элементов по всем разбиениям и выбираются те элементы, которые находились в одном кластере в большей части разбиений. По сути функция устойчивости измеряет, насколько предложенное в ходе передискретизации разбиение на кластеры совпадает с кластерами в исходном наборе данных.

Многие алгоритмы используют параметры, которые определяют разрешение, с которым производится идентификация кластеров. Например, в агломеративных (объединительных) методах кластеризации такие параметры определяют высоту дерева, по которому осуществляется кластеризация. В неиерархических методах такими параметрами может определяться количество допустимых соседей для элемента в кластере.

В данном алгоритме важным является предположение, что параметры алгоритма, влияющие на качество результата, не зависят от размера исходных данных. Суть алгоритма заключается в следующем.

1) Выбираются значения параметров алгоритма кластеризации.

2) Производится кластеризация для всего набора данных.

3) Произвольным образом набор данных разбивается на подмножества.

4) Для каждого подмножества запускается алгоритм кластеризации.

5) На основе шага 1 и 3 с помощью матриц связности вычисляется значение кривой качества.

6) Изменяются параметры алгоритма. Стабильными кластерами считаются те, при которых кривая качества достигает наибольших значений.

Другой робастный алгоритм кластерной устойчивости [11] подсчитывает количество одинаковых меток у элементов, общих для нескольких выборок. Предполагается эффективным для любых алгоритмов кластеризации, однако авторами апробировался на иерархических алгоритмах. В ходе работы алгоритма используются искусственные возмущения исходных данных и строятся дендрограммы. Используются подмножества исходного множества. Для общих точек подмножеств вычисляются схожести кластеризации. Плюсом является то, что для корректной работы не делается никаких предположений о распределении данных или о форме кластеров. Алгоритм позволяет также определить наличие или отсутствие какой-либо структурированности в наборе данных.

5. Заключение

В статье описаны основные алгоритмы устойчивой кластеризации на основе индексов и на основе функций устойчивости. Кластеризация имеет большое количество приложений. Исследование и апробация существующих алгоритмов на реальных данных позволяет узнать их сильные и слабые стороны. На основе этой информации впоследствии могут быть предложены новые алгоритмы, которые будут обладать полезными свойствами. Например, будут требовать небольших вычислительных затрат, будут помехоустойчивыми и способными давать адекватные оценки на множестве данных различной природы.

В качестве возможного решения могут быть предложены, например, алгоритмы многомерной оптимизации, эффективность которых была продемонстрирована в целом ряде прикладных задач.

Поскольку данные методы оптимизации хорошо описаны математически, новые методы также удастся строго описать. Это является важной особенностью, поскольку большинство существующих алгоритмов кластеризации не имеет достаточного математического обоснования.

Были выбраны алгоритмы устойчивой кластеризации на основе индексов и функций устойчивости, поскольку они естественным образом могут быть определены в терминах задач оптимизации. Это позволит получить новые методы решения проблемы кластеризации.

Список литературы

[1] *Calinski R. B., Harabasz J.* A dendrite method for cluster analysis // Communications in Statistics. 3. 1974. pp. 1–27.

[2] *Hartigan J. A.* Clustering Algorithms. — Wiley. 1975.

[3] *Kaufman L., Rousseeuw P.* Finding Groups in Data: An Introduction to Cluster Analysis. — New York: Wiley. 1990.

[4] *Krzanowski W. J., Lai Y. T.* A criterion for determining the number of clusters in a data set // Biometrics. 44. 1985. PP. 23–34.

[5] *Tibshirani R., Walther G., Hastie T.* Estimating the number of clusters in a data set via the gap statistic // Journal of the Royal Statistical Society. Ser. B. 63. 2001. PP. 411–423.

[6] *Volkovich Z., Barzily Z., Morozensky L.* A statistical model of cluster stability // Pattern Recognition. 2008.

[7] *Sugar C., James G.* Finding the number of clusters in a data set: An information theoretic approach // Journal of the American Statistical Association. 98:750–763. 2003.

[8] *Roth V., Lange V., Braun M., Buhmann J.* Stability-based validation of clustering solutions // Neural Computation. 16(6). 2004. PP. 1299–1323.

[9] *Papadimitriou C. H., Steiglitz K.* Combinatorial Optimization, Algorithms and Complexity. Prentice-Hall. Englewood Cliffs. 1982.

[10] *Levine E., Domany E.* Resampling method for unsupervised estimation of cluster validity // Neural Computation. 13. 2001. PP. 2573–2593.

[11] *Ben-Hur A., Guyon I.* Detecting stable clusters using principal component analysis // In Methods in Molecular Biology. Humana press. 2003. PP. 159–182.