УДК 004.021, 004.32.2

РАНДОМИЗИРОВАННЫЙ МЕТОД ОПРЕДЕЛЕНИЯ КОЛИЧЕСТВА КЛАСТЕРОВ НА МНОЖЕСТВЕ ДАННЫХ

Д.С. Шалымов

Рассмотрен новый непараметрический метод устойчивой кластеризации на основе рандомизированного алгоритма стохастической аппроксимации с искусственными возмущениями на входе. Описаны особенности, которые обеспечивают сходимость при почти произвольных (не обязательно гауссовых, но ограниченных по абсолютному значению) помехах. Предлагаемый метод может быть использован в задаче on-line кластеризации для динамически изменяющихся данных. Эффективность демонстрируется примерами.

Ключевые слова: кластеризация, устойчивость кластеризации, рандомизированные алгоритмы, стохастическая аппроксимация, пробное возмущение, on-line алгоритмы.

Введение

Кластеризацией является объединение данных в группы по схожим признакам. Она используется при решении многочисленных задач интеллектуальной обработки данных, в том числе при распознавании образов, машинном обучении, выработке стратегий управления и т.д.

Одной из наиболее сложных задач кластерного анализа является определение количества кластеров, которое получают за счет применения алгоритмов устойчивой кластеризации [1]. Несмотря на кажущееся многообразие, до сих пор не было найдено универсального алгоритма, который был бы эффективным для данных различной природы. Большинство существующих методов основано на индексах, сравнивающих степени «разброса» данных внутри кластеров и между кластерами [2], на расчете значений эвристических характеристик (функций устойчивости), показывающих соответствие назначенных кластеров для выборочных элементов множества [3], на статистиках, определяющих наиболее вероятное решение [4], либо на оценивании плотностей распределений [5]. Чаще всего эти методы либо настроены для конкретных специфических данных, либо требуют определенных предположений о своих параметрах. Кроме того, вычислительная сложность известных алгоритмов кластеризации существенно растет при увеличении мощности исследуемого множества. Также большинство таких алгоритмов недостаточно математически обосновано. Снять эти трудности позволяет использование рандомизированных алгоритмов [6].

В статье предлагается новый непараметрический индексный метод устойчивой кластеризации, основанный на использовании рандомизированного алгоритма типа стохастической аппроксимации (РАСА). Существенная особенность данного алгоритма заключается в том, что при небольших вычислительных затратах на каждой итерации обеспечивается сходимость при почти произвольных помехах. Предлагаемый метод прост в реализации, оказывается эффективным как для искусственно сгенерированных данных с заранее известными свойствами, так и для данных, взятых из реальных практических задач.

Алгоритм устойчивой кластеризации

Устойчивость кластеризации является характеристикой, определяющей различие результирующих разбиений после многократного применения алгоритмов кластеризации. Небольшое расхождение результатов интерпретируется как высокая устойчивость. Количество кластеров, которое максимизирует кластерную устойчивость, может служить хорошей оценкой для реального количества кластеров [7].

В кластерном анализе популярны итеративные методы, которые базируются на априорном знании количества кластеров и некотором выборе первоначального разбиения. Задача нахождения оптимального количества кластеров в общем случае является *NP*-сложной проблемой [8]. Данные часто имеют сложное поведение и не могут быть описаны с помощью аналитических функций. Как правило, это приводит к возникновению шумов и потере информации. Чтобы избежать этого, обычно проводят большее количество итераций, что значительно повышает вычислительную сложность процесса кластеризации. Известно, что в других приложениях снять эти трудности удается при использовании рандомизированных алгоритмов, сложность которых, как правило, не зависит от роста размерности исходной задачи [6].

Известен эффективный метод устойчивой кластеризации, предложенный исследователями Sugar и James [9], основанный на использовании внутренних дисперсий кластеров, называемых «искажениями». С помощью алгоритма *k-means* [10] множество разбивается на кластеры, далее строится кривая зависимости минимального «искажения» от текущего количества кластеров. Теоретически и эмпирически доказывается, что при определенном выборе параметра Y (степень трансформации) вышеупомянутая кривая будет иметь резкий скачок в том месте, которое соответствует настоящему количеству кластеров.

Основная процедура метода состоит из следующих шагов.

1. Запускается *k-means* алгоритм для K кластеров и определяется соответствующее «искажение» d_k. Для различных значений K строится набор d_k.

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

3. Вычисляются скачки по формуле $J_K = d_K^{-Y} - d_{K-1}^{-Y}$.

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

Вместо алгоритма *k-means* предлагается использовать альтернативный алгоритм PACA с двумя измерениями [12], который оказывается эффективным в многомерных задачах с большим количеством параметров и наличием разнообразных помех, а также не обладает, в отличие от *k-means*, свойством вырожденности, когда на какой-либо итерации получается пустой кластер.

На рис. 1 изображены зависимости «искажений» модифицированного метода от количества кластеров. В примере для входных данных было использовано гауссово распределение, однако метод успешно зарекомендовал себя также на негауссовых данных.

Единственным параметром, зависящим в общем случае от данных, является степень трансформации. Есть возможность отказаться от него, если свести вышеописанную задачу устойчивой кластеризации к оптимизационной задаче. Для этого исследуется кривая на рис. 1, с,, которая приближается гладкой функцией и имеет одну ярко выраженную экстремальную точку. Для поиска этой точки в условиях помех можно воспользоваться, например, классическим методом конечно-разностной стохастической аппроксимации [11]. Еще одним способом отказа от параметра является использование степени трансформации, обратно пропорциональной текущему количеству кластеров.

Учитывая тот факт, что кривая на рис. 1, c, аппроксимируется прямой на участках до и после истинного значения количества кластеров, можно использовать метод ломаной [9]. Тогда за результирующее значение количества кластеров можно принять то, которое минимизирует квадратичное отклонение кривой от двух прямых ломаной.

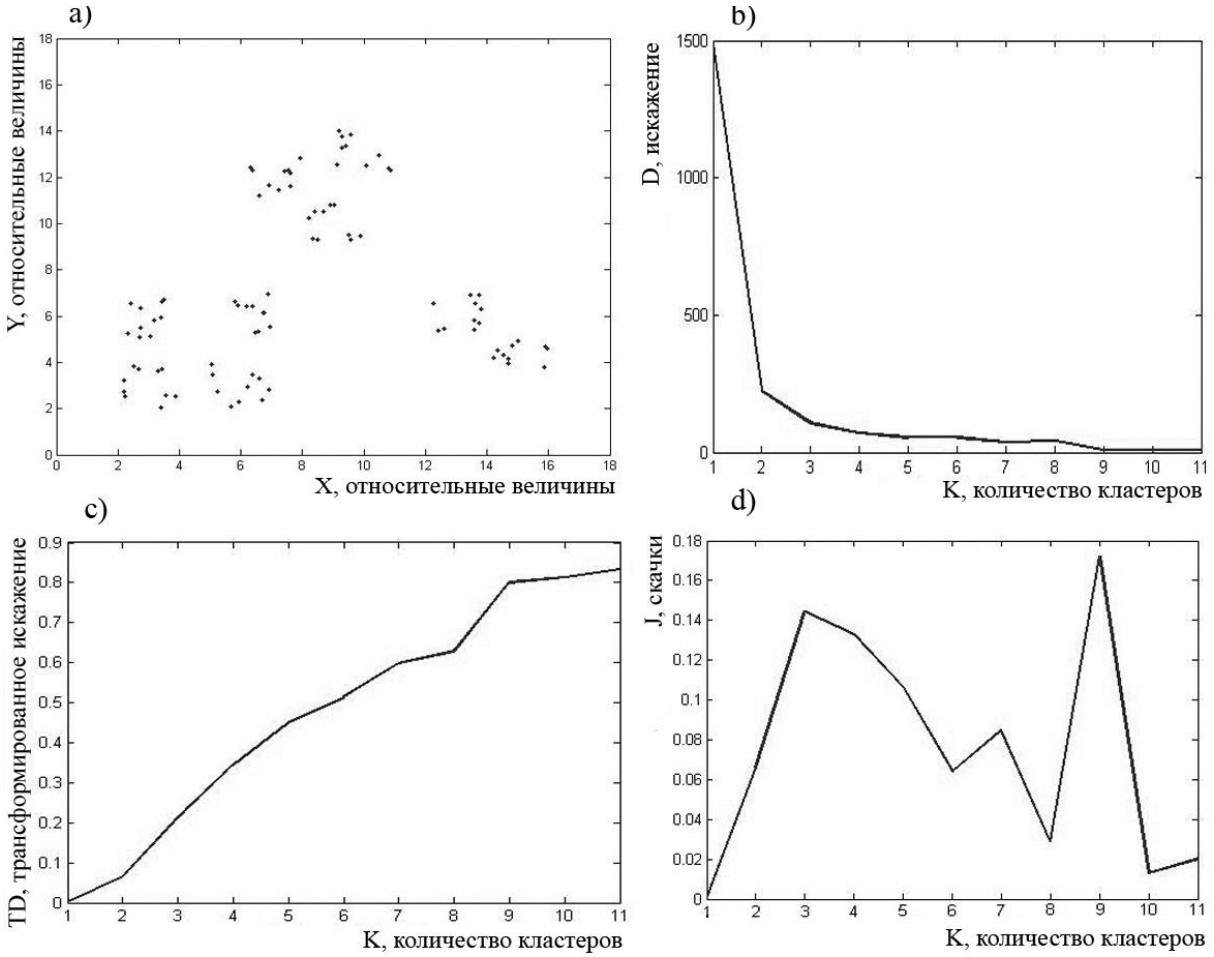

Рис. 1. Иллюстрация работы предлагаемого метода устойчивой кластеризации: a) входные двухмерные данные, представляющие собой набор из девяти гауссовских кластера; b) кривая зависимости внутренней дисперсии от количества кластеров; c) кривая преобразованных внутренних дисперсий; d) кривая зависимости скачков от количества кластеров

Предлагаемый метод оказывается также эффективным в задаче определения структуры множества. Анализируя кривую скачков на рис. 1, d, можно увидеть, что в качестве основного ответа метод предлагает количество кластеров, равное девяти, а в качестве альтернативного – трем.

Автоматическая классификация входных данных

Для определения «правильного» количества кластеров на множестве данных многократно запускаются алгоритмы автоматической классификации с заранее известным количеством классов. С содержательной точки зрения смысл автоматической классификации состоит в построении правила, сопоставляющего каждой точке x множества X некоторый образ (класс). При этом должна быть задана функция расстояния между объектами $\rho(x, x')$.

Всякий способ классификации связан с потерями, которые обычно характеризуются с помощью штрафных функций (стоимости) $q^k(x, \eta)$, $k = 1, 2, \ldots, l$, η – набор век-

торов, характеризующий центры классов. В типичных случаях, когда X – вещественное векторное пространство, значения штрафных функций $q^k(x, \eta)$ возрастают при удалении x от центра соответствующего образа (класса). Геометрический смысл задачи автоматической классификации заключается в следующем. Допустим, что в системе всего l классов $\eta = (\Theta^1, \Theta^2, \ldots, \Theta^l)$, а штрафные функции имеют похожий друг на друга вид $q^k(x,\eta) = \left\| x - \theta^k \right\|^2$, $k = 1,2,...,l$. Рассмотрим разбиение множества X на l классов $X^1(\eta), X^2(\eta),..., X^l(\eta)$ по правилу: к множеству $X^k(\eta)$ относятся все точки x, которые находятся к центру θ^k ближе, чем к любому другому. Интеграл

$$\int\limits_{X^k(\eta)} \left\| x - \theta^k \right\|^2, \ 1, 2, \ldots, l,$$

определяет рассеяние точек x в множестве $X^k(\eta)$, а функционал среднего риска есть

$$F(\eta) = \sum_{k=1}^{l} \int_{X^k(\eta)} \left\| x - \theta^k \right\|^2 P(dx).$$

Задача автоматической классификации состоит в определении набора центров $\{\theta_*^k, k = 1,2,...,l\}$, при которых суммарное рассеивание минимально. Решение не обязательно должно быть единственным, поскольку при перестановке местами векторов внутри набора $\{\theta_*^k, k = 1,2,...,l\}$ значение определенного выше функционала среднего риска не изменяется.

На практике функции $q^k(\cdot,\cdot), k = 1,2,...,l$ не всегда заданы аналитически, но их значения доступны измерению (может быть, с помехами v^k):

$y^k(x,\eta) = q^k(x,\eta) + v^k, k = 1,2,...,l.$

Если функционал $F(\eta)$ дифференцируем, то искомый набор центров η_* должен удовлетворять уравнению $\nabla F(\eta_*) = 0$. Но при решении рассматриваемой задачи нельзя воспользоваться традиционными градиентными методами, так как из его вида понятно, что он не везде дифференцируем, и, кроме того, не всегда возможно прямое вычисление $\nabla F(\eta)$. Для аппроксимации градиента функционала среднего риска будем использовать РАСА с двумя измерениями на входе [12]. Известно, что при выполнении определенных условий РАСА по сравнению с классическим конечно-разностным алгоритмом стохастической аппроксимации (процедура Кифера–Вольфовица) [11] обеспечивает одинаковую точность за одно и то же количество итераций. При этом РАСА требуется в M (где M – размерность пространства) раз меньше измерений функции, что обеспечивает большой выигрыш в вычислительной сложности. РАСА основан на использовании наблюдаемой последовательности серии случайных независимых друг от друга векторов $\Delta_j \in R^m$, $j = 1,2,...$, называемых пробным одновременным возмущением и составленных из независимых бернуллиевских случайных величин. В [12] доказывается, что подобные алгоритмы сходятся к оптимальному набору центров классов η_* при почти произвольных помехах.

Практическое применение

Существует пять наиболее известных алгоритмов устойчивой кластеризации. Метод *CH* (Calinski–Harabasz) [5] выбирает количество кластеров, максимизирующее функцию $CH(K) = \dfrac{B(K)/(K-1)}{W(K)/(n-K)}$, где $B(K)$ и $W(K)$ − средние межкластерные и внутри-

кластерные расстояния. Подход *KL* (Krzanowski and Lai) [13] максимизирует функцию $KL(K) = \left| \frac{DIFF(K)}{DIFF(K+1)} \right|$, где $DIFF(K) = (K-1)^{2/p} W(K-1) - K^{2/p} W(K)$. Hartigan [4] предлагает выбирать наименьшее значение *K,* при котором значение функции $H(K) = (n-K-1)\left[\frac{W(K)}{W(K+1)} - 1 \right]$ меньше либо равно 10. Silhouette [1] измеряет, насколько хорошо была кластеризована *i*-я точка, для чего определяется функция $s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$, 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 $s(i)$. Метод GAP [14] использует статистику расхождений, вычисляя функцию $GAP(K) = \frac{1}{B} \sum_b \log(W_b^*(K)) - \log(W(K))$. Метод использует *B* различных унифицированных множеств. Требуется подобрать *K,* максимизирующее $GAP(K)$.

Предлагаемый метод на основе PACA был сравнен с вышеописанными алгоритмами для данных, сгенерированных по трем различным сценариям. Результаты представлены в таблице.

Данные	**Метод**	**Оценка количества кластеров**									
		1	2	3	4	5	6	7	8	9	10
I сценарий	*CH*	0	96	4	0	0	0	0	0	0	0
5 гауссовских кластера,	*KL*	0	0	0	0	98	0	1	1	0	0
размерность пространства	*Hartigan*	0	0	0	0	100	0	0	0	0	0
равна 10	*Silhouette*	0	100	0	0	0	0	0	0	0	0
	GAP	0	0	0	0	100	0	0	0	0	0
	PACA	0	0	0	0	100	0	0	0	0	0
II сценарий	*CH*	0	0	0	83	5	5	3	0	1	1
4 гауссовских кластера с	*KL*	0	0	0	76	7	2	3	8	4	0
различными ковариациями,	*Hartigan*	0	0	0	0	0	0	8	20	23	49
размерность пространства	*Silhouette*	0	34	0	65	1	0	0	0	0	0
равна 10	*GAP*	0	20	0	78	2	0	0	0	0	0
	PACA	0	0	0	100	0	0	0	0	0	0
III сценарий	*CH*	0	0	0	22	11	19	10	6	15	17
4 кластера с экспоненциальным	*KL*	0	0	0	71	17	4	3	0	5	0
распределением,	*Hartigan*	0	0	0	0	0	2	7	9	17	65
размерность пространства	*Silhouette*	0	0	0	60	30	8	1	1	0	0
равна 10	*GAP*	85	9	0	6	2	0	0	0	0	0
	PACA	0	0	0	99	1	0	0	0	0	0

Таблица. Применение алгоритмов устойчивой кластеризации к искусственным данным, сгенерированным по трем различным сценариям

Для каждого сценария было сгенерировано по 100 тестовых множеств. Предлагаемый метод оказывается помехоустойчивым. Каждый из классических подходов, в отличие от PACA, для данных второго и третьего сценариев дал неверные результаты как минимум в двух случаях.

Метод PACA был успешно применен для данных, ставших классическими для тестирования алгоритмов кластеризации, а также для данных, количество кластеров в которых изменялось во времени. С помощью предложенного метода можно определить не только количество кластеров и структуру данных, но также оценки координат центров кластеров.

Заключение

Рассмотренный непараметрический индексный метод устойчивой кластеризации на основе РАСА оказывается эффективным как на искусственных, так и на реальных данных. Метод прост в применении. Кроме того, в отличие от большинства существующих алгоритмов кластеризации, он может быть строго описан математически. РАСА остается устойчивым в условиях почти произвольных помех. При этом с ростом размерности фазового пространства его сложность не возрастает.

В качестве дальнейшего исследования предполагается рассмотреть альтернативные способы определения количества кластеров на основе анализа кривой «искажений», а также апробировать существующие модификации РАСА с одним и двумя измерениями функции стоимости на каждой итерации.

Литература

1. Kaufman L. and Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. – New York: Wiley, 2005. – 368 с.
2. Wishart D. Mode analysis: A generalisation of nearest neighbour which reduces chaining effects. // Numerical Taxonomy. – 1969. – С. 282–311.
3. Levine E. and Domany E. Resampling method for unsupervised estimation of cluster validity // Neural Computation – 2001. – № 13. – С. 2573–2593.
4. Hartigan J.A. Clustering Algorithms. – New York: John Wiley, 1975. – 351 с.
5. Calinski R. and Harabasz J. A dendrite method for cluster analysis // Commun Statistics. – 1974. – № 3. – С. 1–27.
6. Граничин О.Н., Поляк Б.Т. Рандомизированные алгоритмы оптимизации и оценивания при почти произвольных помехах. – М., Наука, 2003. – 291 с.
7. Volkovich Z., Barzily Z., Morozensky L. A statistical model of cluster stability // Pattern Recognition. – 2008. – № 41. – С. 2174 – 2188.
8. Hansen P. and Mladenovic N. J-means: a new local search heuristic for minimum sum-of-squares clustering // PatternRecognition. – 2002. – № 34(2). – С. 405–413.
9. Sugar C. and James G. Finding the number of clusters in a data set : An information theoretic approach. // Journal of the American Statistical Association. – 2003. – № 98. – С. 750 –763.
10. Hartigan J.A., Wong M.A. A K-Means Clustering Algorithm // Applied Statistics. – 1979. – № 28. – С. 100 – 108.
11. Kushner H.J., Yin G.G. Stochastic Approximation Algorithms and Applications. – New York, Springer-Verlag, 2003. – 474 с.
12. Граничин О.Н., Измакова О.А. Рандомизированный алгоритм стохастической аппроксимации в задаче самообучения // Автоматика и телемеханика. – 2005. – № 8. – С. 52–63.
13. Krzanowski W.J., Lai Y.T. A criterion for determining the number of clusters in a data set // Biometrics. – 1985. – № 44. – С. 23–34.
14. 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. – 2001. – № 63. – С. 411–423.

Шалымов Дмитрий Сергеевич – Санкт-Петербургский государственный университет, аспирант, shalydim@mail.ru