УДК 004.023

БОЛДАК А.А.,
СУХАРЕВ Д.Л.

ОПРЕДЕЛЕНИЕ КОЛИЧЕСТВА КЛАСТЕРОВ В СТАТИСТИЧЕСКИХ ДАННЫХ

В работе разработаны и экспериментально изучены три эвристических алгоритма автоматического определения количества кластеров выборок данных, использование которых позволяет усовершенствовать известные алгоритмы кластеризации данных. Особенностью предложенных алгоритмов является отсутствие необходимости многократного решения задачи кластеризации для разного количества кластеров, с последующим анализом качества кластерной структуры.

In the work three heuristic algorithms for automatic determination of the clusters amount of data samples were developed and experimentally studied. Usage of them can improve well-known clustering algorithms. Feature of the proposed algorithms is that there is no need to execute multiple calculations of clustering task for different clusters amount, with subsequent analysis of cluster structure quality.

Введение

Алгоритмы кластеризации являются одним из инструментов, которые с успехом используются для решения задач, связанных с интеллектуальным анализом данных (Data Mining) [1], обработкой изображений [2], рас познаванием образов [2], группированием результатов поиска [3], сжатием данных [3] и т.п. Особенность известных методов иерархической и разбивающей кластеризации [2] состоит в том, что для них количество кластеров является параметром, автоматическое определение которого может рассматриваться как самостоятельная задача. Известные алгоритмы её решения (Gap подход [4], G-Means [5], X-Means [6]) предполагают повторное выполнение кластеризации с последующей оценкой качества полученной кластерной структуры. Очевидно, что применение таких методов требует значительных вычислительных затрат для решения этой задачи.

Цель работы состоит в снижении вычислительных затрат, необходимых для автоматического определения количества кластеров за счет разработки и использования алгоритмов решения этой задачи, отличительной особенностью которых является отсутствие необходимости повторного выполнения кластеризации данных.

1. Постановка задачи

Постановка задачи кластеризации связана с определением метрического пространства

$$(O, d : O \times O \to R),$$

где d – метрика, O – множество объектов кластеризации, $O = \{O_i\},\ i = \overline{1, |O|}$.

Кластеризация объектов множества O представляет собой отображение вида:

$$f : O_i \to C_j, i = \overline{1, |O|}, j = \overline{1, k},$$

где C – множество кластеров, k – их количество, $k = \overline{1, |O|}$.

Для каждого кластера C_j с множеством элементов $O_j \subseteq O$ можно определить центроид, т.е. такой объект c_j, для которого выполняется условие:

$$c_j : \sum_{i=1}^{|O|} d(c_j, O_{ji})^2 \to \min.$$

Как правило, формальная кластеризация предполагает, что сведения обо всех объектах множества O заданы с помощью количественных оценок их свойств (показателей) в виде таблицы «объект-свойство» [7]:

$$X = (x_{ij})_{i=1, j=1}^{|O|, m}, x_{ij} \in R,$$

в которой строка X_i соответствует множеству значений показателей для объекта O_i, столбец X^j – множеству значений j-го показателя для всех объектов множества O, который рассматривается как статистическая переменная [7].

В этом случае метрика d выражается как метрика в пространстве R^m и интерпретируется как оценка близости объектов в R^m. В один кластер включаются близкие друг к другу по метрике d объекты.

Пусть качество результата работы алгоритма кластеризации f оценивается количественно с помощью критерия $K(O, k)$, O – множество объектов, k – количество кластеров. Тогда задача определения параметра k может быть поставле-

на как задача поиска аргумента максимизации (argmax) или минимизации (argmin) этого критерия:

$$k = \arg\max(K(X,k)),\ k = \overline{1,|O|},$$
$$k = \arg\min(K(X,k)),\ k = \overline{1,|O|}.$$

Вид и содержание критерия *K*, и связанная с ним задача оптимизации определяются семантикой конкретного прикладного применения метода кластеризации.

2. Эвристические алгоритмы определения количества кластеров

Предлагаемые алгоритмы определения количества кластеров основаны на предположении, что кластерная структура данных характеризуется неравномерностью плотности распределения значений X^j.

Функция плотности распределения значений статистической переменной $p(X^j,n)$, где n – количество уровней дискретизации, соответствующих интервалам группирования значений этой переменной [8], задаёт вероятность событий вида $X^j = x_{ij}$, $i = \overline{1,n}$. Очевидно, что равномерное распределение значений статистической переменной (X_U^j) соответствует отсутствию кластерной структуры в данных. Это означает, что степень ее проявления можно оценить с помощью количественной характеристики отклонения функции плотности распределения значений X^j от функции плотности равномерного распределения X_U^j в области значений X^j.

Таким образом, разработка алгоритма определения количества кластеров для одной статистической переменной требует определения меры μ в пространстве функций плотности распределения вероятностей и такого количества интервалов группирования данных, на котором расстояние между $p(X^j,n)$ и $p(X_U^j,n)$, оцененное по мере μ, будет максимальным.

Информационная энтропия. Для дискретной случайной величины *X*, принимающей значения $\{x_i : i = \overline{1,n}\}$, шенноновская энтропия, мера неопределенности, представляется как [9]:

$$H(X^j) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i).$$

Для существования кластерной структуры вероятностное распределение данных X^j должно отличаться от равномерного. Поэтому в качестве меры μ можно использовать понятие эффективности алфавита [10]:

$$E(X^j) = -\sum_{i=1}^{n} \frac{p(x_i)\log_b p(x_i)}{\log_b(n)}, \qquad (1)$$

где числителем является энтропия для X^j, знаменателем – энтропия для X_U^j. Это значение всегда меньше или равно единицы. Чем оно меньше, тем более предсказуемым является исходное распределение по отношению к равномерному, и тем легче выделяются интервалы с большей вероятностью попадания в них. В случае кластеризации такие интервалы будут соответствовать кластерам, то есть интервалы группирования можно рассматривать как алфавит мощности *n*.

При решении задачи кластеризации в правой части формулы (1) известны значения всех операндов, кроме *n*. Тогда формула (1) может рассматриваться как выражение для вычисления критерия $K(X^j,n)$, а задача поиска *n* – как

$$n = \arg\min(K(X^j,n)).$$

Если начальным количеством интервалов группирования есть значение $n=N_{max}$, являющееся степенью двойки

$$N_{\max} = 2^p,\ p \in N,$$

то попарное суммирование значений $p(x_i),\ i = \overline{1,N_{\max}}$ даёт значения $p(x_i),\ i = \overline{1,N_{\max}/2}$ для $N_{\max}/2$ интервалов и т.д. Это позволит уменьшить число обращений к исследуемым объектам.

Дивергенция Кульбака-Лейблера. Альтернативным методом определения количества интервалов группирования является дивергенция Кульбака-Лейблера [11][12][13], известная также как относительная энтропия:

$$D_{KL}(P \| Q) = \sum_{i=1}^{n} P(x_i)\log\frac{P(x_i)}{Q(x_i)}.$$

В случае кластеризации наибольший интерес она представляет как расстояние между выборками *P* и *Q*. Однако она не является полноценной метрикой т.к. не удовлетворяет условию симметричности. Поэтому используют её симметризированную форму – дивергенцию Дженсона-Шеннона (JSD) [14]:

$$D_{JS}(P \| Q) = \frac{1}{2} D_{KL}(P \| M) + \frac{1}{2} D_{KL}(Q \| M), \quad (2)$$

где M – средняя плотность распределения для P и Q:

$$M = \frac{1}{2}(P + Q). \quad (3)$$

Пусть в формулах (2),(3) P – распределение исходных данных X^j, Q – равномерное распределение X_U^j. Тогда задача нахождения интервалов группирования может быть поставлена как

$$n = \arg\max(K(P,Q,n)), \; n = \overline{1, N_{\max}},$$

но дивергенция здесь не может использоваться как критерий K, т.к. возрастает с ростом количества интервалов группирования. Однако она поддается аппроксимации функцией $f(n) = a \log n$. Тогда критерий K имеет следующий вид:

$$K(P,Q,n) = D_{JS}(P \| Q) - a \log(n).$$

Для уменьшения количества вычислений используются те же оптимизации, что и в предыдущем алгоритме, основанном на энтропии.

Деление интервалов группирования до образования локального минимума (ДИГОЛМ). С одной стороны, поиск центроидов связан с нахождением плотных интервалов группирования, в которых локализованы эти центроиды. С другой стороны, поиск границ связан с нахождением разреженных интервалов. Таким образом, задача связана с поиском экстремумов: локальных минимумов x_{min} и максимумов x_{max} функции вероятности $p(x_i)$:

$$\begin{aligned} &p(x_{\min}) \le p(x_i) \; при \; |i - \min| \le \varepsilon, \; i = \overline{1,n}, \\ &p(x_{\max}) \ge p(x_i) \; при \; |i - \max| \le \varepsilon, \; i = \overline{1,n}, \end{aligned} \quad (4)$$

где ε – единица, т.к. номера интервалов группирования являются дискретными величинами.

При этом, также как и в предыдущем случае, значение n выбирается среди степеней двойки. Это позволяет сократить объем необходимых вычислений для решения задачи (4).

Поиск центроидов. Задача поиска кластеров может решаться, как задача локализации плотных областей в многопараметрическом пространстве. Центры этих областей соответствуют центроидам предполагаемых кластеров.

Когда данные по каждому показателю сгруппированы в интервалы и при этом наблюдается неравномерность распределения плотности, можно построить *многомерные интервалы группирования (МИГи)*.

Каждому МИГу соответствует один интервал группирования для каждого показателя. Тогда количество МИГов N_B можно определить из формулы:

$$N_B = \prod_{i=1}^{n} l_i,$$

где l_i – количество интервалов группирования i-го показателя, n – размерность пространства.

Затем рассчитывается функция плотности объектов, попадающих в каждый МИГ. Объект x_i считается принадлежащим МИГу когда он попадает во все интервалы I_i, включенные в этот МИГ, то есть

$$\bigcap_{i=1}^{n} (x_i \in I_i) = 1,$$

где n – размерность пространства.

Разбиение пространства на многомерные интервалы группирования и оценка их плотности позволяет определить приблизительное положение центроидов, которые могут быть использованы как начальные центроиды алгоритма k-means. Это дает возможность уменьшить количество итераций этого алгоритма.

Для определения центроидов вычисляется ожидаемое значение функций плотности МИГов

$$M_x = \frac{1}{N_B} \sum_{i=1}^{N_B} p_i.$$

Если значение функции плотности МИГа превышает M_x ($p_j > M_x$), то центр этого МИГа признается центроидом. Количество кластеров определяется количеством центроидов.

Оценка эффективности. В разработанных алгоритмах кластеризации данных, которые автоматически определяют количество кластеров, применяются эвристики поиска количества интервалов группирования, поиска центроидов, а также методы кластеризации. К реализованным алгоритмам поиска количества интервалов группирования относятся эффективность алфавита теории энтропии, дивергенция Дженсена-Шеннона и ДИГОЛМ. Алгоритмы поиска центроидов делятся на аналитические и случайные. В качестве алгоритма кластеризации используется метод k-means. Таким образом, было разработано 6 алгоритмов кластеризации, способных автоматически определять количество кластеров.

Были созданы средства генерации экспериментальных выборок данных, которые возвращают выборки разного вида, в зависимости от

количества кластеров, их плотности и разделенности.

Для статистической достоверности оценка эффективности алгоритмов кластеризации была выполнена для 4000 выборок данных разного вида. В таблице 1 приведены усредненные результаты этих оценок.

Для оценки качества кластерной структуры используются индекс Данна [15] и среднее значение силуэта (average silhouette) [16]. Они имеют достаточно высокую точность и напрямую не зависят от количества кластеров. Кроме того, среднее значение силуэта является нормированной величиной и устойчиво к вбросам. Для этих критериев большее значение свидетельствует о более высоком качестве кластерной структуры.

Оценка производительности алгоритмов кластеризации выполняется на основе количества итераций и длительности их выполнения в миллисекундах. Измерения проводились на ПК с процессором Intel Core 2 Duo E8500 под управлением ОС Ubuntu 10.04 x64.

Табл. 1. Усредненные результаты измерений производительности применения различных эвристик

Определение кол-ва интервалов	Определение начальных центроидов	Кол-во итераций	Длительность, мс	Кол-во кластеров	Індекс Данна	Среднее значение силуэта
Энтропия	Аналитическое	6.696	**302.05**	11.918	**2.307**	0.750
ДИГОЛМ	Аналитическое	**6.632**	316.52	6.210	1.639	0.697
JSD	Аналитическое	6.736	337.74	12	2.297	**0.763**
Энтропия	Случайное	27.473	1048.5	11.918	1.335	0.297
ДИГОЛМ	Случайное	11.775	447.82	6.215	1.375	0.626
JSD	Случайное	27.117	1079.5	12	1.361	0.338

В соответствии с результатами, приведенными в таблице 1, можно сделать вывод, что в общем случае алгоритм на основе метода ДИГОЛМ имеет наименьшее среднее количество итераций, алгоритм на основе метода энтропии – наименьшее время выполнения и наибольший индекс Данна, а алгоритм на основе JSD – наибольшее значение среднего силуэта. Среднее значение количества кластеров, реально возвращаемых генераторами экспериментальных выборок данных, равнялось 13. Самые близкие к этому значению результаты показали алгоритмы, использующие методы энтропии и JSD. Алгоритмы на основе метода ДИГОЛМ плохо различают большое количество (больше шести) кластеров. Алгоритмы со случайным определением начальных центроидов значительно уступают тройке алгоритмов с их аналитическим определением.

Заключение

Результатом работы являются три эвристических алгоритма автоматического определения количества кластеров, использование которых позволяет усовершенствовать известные алгоритмы кластеризации данных, в частности алгоритм k-средних и его модификации. Эти усовершенствования касаются как качества кластерной структуры, так и времени выполнения.

Основой разработанных алгоритмов есть то, что наличие кластерной структуры может количественно оцениваться отклонениями функции плотности распределения статистической переменной от функции плотности равномерного распределения. Для оценки этих отклонений используются эффективность алфавита теории энтропии, дивергенция Дженсена-Шенонна и ДИГОЛМ. Т.к. они оперируют с функциями плотности, которые являются дискретными и могут быть получены за один проход данных, то такие алгоритмы имеют линейную вычислительную сложность.

В результате анализа экспериментальных данных, полученных с использованием разработанных программных средств, можно сделать вывод, что в общем случае, когда вид выборок данных неизвестен, алгоритмы с использованием JSD и эффективности алфавита теории энтропии дают наилучшие результаты.

Использование описанных алгоритмов позволяет значительно снизить вычислительные затраты, необходимые для автоматического определения количества кластеров сравнительно с существующими алгоритмами, требующими множественного выполнения кластеризации данных.

Список литературы

1. Usama F., Piatetsky-Shapiro G., Smyth P. (1996). From Data Mining to Knowledge Discovery in Databases // AI MAGAZINE. – 1996. – pp. 37-54.
2. Jain A.K. and Dubes R.C. Algorithms for Clustering Data // Prentice Hall. – 1988. – 320 c.
3. Tan P.N., Steinbach M., Kumar V. Introduction to Data Mining // Addison-Wesley. – 2005. – 769 c.
4. Tibshirani R., Walther G., Hastie T. Estimating the number of clusters in a dataset via the Gap statistic. // Technical Report. Stanford. – 2000. – pp. 412-423.
5. Hamerly G., Elkan C. Learning the k in k-means. // NIPS. – 2003.
6. Pelleg D., Moore A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. // Morgan Kaufmann. – 2000.
7. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика. 1985. – 480 с.
8. Кудрявцев Л. Д. Курс математического анализа. — 5-е изд. — М.: «Дрофа», 2003. — Т. 1. — 704 с.
9. Shannon C.E. A Mathematical Theory of Communication // Bell System Tech. – 1948. – vol. 27, pp. 379-423, 623-656.
10. Potamites P. Selective Pressures on Symbolic Systems. [Електроний ресурс] – January 28, 2010. – Режим доступу: http://www-scf.usc.edu/~potamite/philqual2.pdf
11. Kullback S., Leibler R.A. On Information and Sufficiency. Annals of Mathematical Statistics. – 1951. – 22 (1): pp. 79–86.
12. Kullback S. Information theory and statistics. // John Wiley and Sons, NY. – 1959. – 416 c.
13. Kullback S. Letter to the Editor: The Kullback-Leibler distance // The American Statistician. – 1987. – 41(4): pp. 340–341.
14. Fuglede B., Topsoe F. Jensen-Shannon divergence and hilbert space embedding. // University of Copenhagen, Department of Mathematics. – 2004.
15. Dunn J. Well separated clusters and optimal fuzzy partitions // Journal of Cybernetics. – 1974. – 4, pp. 95-104.
16. Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. – 1987. – 20, pp. 53-65.