УДК 510.63; 519.68
DOI: 10.15827/0236-235X.120.643-646
Дата подачи статьи: 31.03.17
2017. Т. 30. № 4. С. 643–646

ОПТИМАЛЬНАЯ ЭНТРОПИЙНАЯ КЛАСТЕРИЗАЦИЯ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Б.Г. Аскерова, к.т.н., доцент, Bahar287@mail.ru
(Азербайджанский государственный университет нефти и промышленности, просп. Азадлыг, 20, г. Баку, AZ1010, Азербайджан)

В данной работе исследована возможность разработки нового метода кластеризации данных в информационных системах. Кластеризация – это процесс нахождения возможных групп в заданном множестве с учетом признаков схожести или различия элементов этого множества. Существующий метод энтропийной кластеризации представляет собой информационно-теоретический подход к задаче кластеризации. В статье предлагается метод оптимальной энтропийной кластеризации высокоразмерных данных в информационных системах, который базируется на энтропийном подходе к выбору состояния элементов сообщений. Дано его математическое обоснование.

Разработанный метод оптимальной энтропийной кластеризации базируется на известном принципе «малая величина энтропии соответствует большому количеству информационного содержания» и позволяет формировать режим не только оптимальной кластеризации, но и сокращения признакового пространства.

Предложены методики вычисления степени оптимальности проведенной кластеризации, а также сокращения признакового пространства высокоразмерных данных при их первичной обработке.

Ключевые слова: *кластеризация, оптимизация, высокоразмерные данные, информационные системы, энтропия.*

Кластеризация объектов сложной природы с высокой размерностью признакового пространства является актуальной задачей во многих областях научных исследований [1–5]. Среди методов сокращения размерности пространства признаков в качестве основных выделяются методы главных компонент и нормализации [1]. Первый из указанных методов чересчур чувствителен к методам предобработки данных, а второй требует обоснованного выбора метода нормализации высокоразмерных данных. Согласно [1], важной и актуальной задачей является разработка эффективных методов предобработки данных и сокращения размерности признакового пространства без существенной потери информации.

Кластеризация – это процесс нахождения различных групп в заданном множестве с учетом признаков схожести или различия элементов этого множества [3]. При этом в качестве меры схожести или различия используют метрику в виде расстояния между векторами x и y:

$$d(x,y)=\left(\sum_{i=1}^{l} w_i \left|x_i-y_i\right|^p\right)^{1/p},$$

где x_i, y_i – i-я координата векторов x, y, $i=\overline{(1,l)}$; w_i – i-й весовой коэффициент. При $w=1$ выражение представляет собой расстояние Минковского порядка p, при $p=2$ получаем расстояние Эвклида, при $p=1$ – расстояние Манхеттена, при $l=\infty$ – расстояние Чебышева.

В работе [4] кластеризация представляется как первичный анализ большого объема данных высокой размерности при отсутствии априорных знаний о них.

В методе SEC [4] кластеры организуются по признаку эффективности сжатия бинарных векторов при использовании отдельного кодера для каждой группы.

В работе [5] решена задача подбора признака для кластеризации на основе количественной характеристики, вырабатываемой с помощью четырех предложенных алгоритмов классификации. При этом подбор признака осуществляется методом, позволяющим определить подмножества исходных признаков, имеющих одинаковую смысловую информацию в отношении БД.

Качественная предобработка признаков объекта может быть осуществлена по методу энтропии, используя условие минимума энтропии Шеннона, что соответствует максимальной информации об изучаемых объектах [1, 6, 7]. Что касается практической пользы кластеризации, то, например, согласно [2, 8], она помогает идентифицировать группы генов, имеющих сходные образы экспрессии при различных условиях. Такие гены типичным образом вовлечены в выполнение связанных функций. В работе [2] предлагается информационно-теоретический подход к кластеризации данных экспрессии генов. Известно, что энтропия является мерой информации и неопределенности случайного переменного. Следовательно, критерием кластеризации может стать условие достижения минимума энтропии. При использовании критерия минимальной энтропии проблема кластеризации имеет две субпроблемы: оценка a posteriori вероятности и минимизация энтропии. При этом, так как энтропия является мерой беспорядка в системе, каждый кластер должен иметь минимальную энтропию. Другими словами, данные в одном и том же кластере должны иметь схожие числовые характеристики. Существующий метод энтропийной кластеризации представляет собой информационно-теоретический подход к задаче кластеризации. Например, в простом случае каждый отдельный кластер должен содержать объекты с одинаковой величиной энтропии [9, 10], однако су-

ществующий метод энтропийной кластеризации может быть подкреплен известным принципом «малая величина энтропии соответствует большому количеству информационного содержания». Учет данного принципа позволяет по-новому подойти к методу энтропийной кластеризации и в сущности разработать новый метод оптимальной энтропийной кластеризации.

Предлагаемый метод оптимальной энтропийной кластеризации

Предлагаемый в настоящей статье метод кластеризации базируется на энтропийном подходе к выбору состояния элементов сообщения. Допустим, имеется множество X независимых элементов x_i, $i = \overline{(1, m)}$, то есть $X = \{x_i\}$.

При этом формируются кластеры $K(P_j)$, $j = \overline{(1, m)}$.

Порядок формирования кластеров такой, что в кластер $K(P_j)$ включаются элементы, имеющие P_j состояний.

Если обозначить количество элементов в каждом j-м кластере как n_j, то общее количество информации, содержащейся во всех элементах одного j-го кластера, вычислим как

$$M_j = n_j \cdot \log_2 P_j . \quad (1)$$

Суммируя (1) по всем j, получим

$$\sum_{j=1}^{m} M_j = \sum_{j=1}^{m} n_j \cdot \log_2 P_j . \quad (2)$$

Далее допустим существование функциональной связи между переменными n_j и P_j, то есть

$$P_j = \varphi(n_j) . \quad (3)$$

Также допускается существование определенного ограничения на сумму $\sum_{j=1}^{m} \varphi(n_j)$, то есть

$$\sum_{j=1}^{m} P_j = \sum_{j=1}^{m} \varphi(n_j) = C, \text{ где } C = \text{const}. \quad (4)$$

Кластеризацию элементов x_i по предлагаемому информационно-вариационному критерию будем считать оптимальной, если при вычисленной оптимальной функции $\varphi(n)_{opt}$ общее количество информации, определяемое выражением

$$F_1 = \sum_{j=1}^{m} n_j \log_2 \varphi(n_j)_{opt} + \lambda \cdot \sum_{j=1}^{m} \varphi(n_j)_{opt} , \quad (5)$$

где λ – множитель Лагранжа, с учетом условия (4) достигает максимальной величины.

Для оценки степени оптимальности реальной кластеризации введем на рассмотрение коэффициент оптимальности, определяемый как

$$\chi = \frac{\sum_{j=1}^{m} n_j \log_2 \varphi(n_j)_{real} + \lambda \cdot \sum_{j=1}^{m} \varphi(n_j)_{real}}{\sum_{j=1}^{m} n_j \log_2 \varphi(n_j)_{opt} + \lambda \sum_{j=1}^{m} \varphi(n_j)_{opt}} , \quad (6)$$

где $\varphi(n_j)_{real}$ – реальная функция зависимости P_i от количества элементов n_j в кластере P_j.

Покажем порядок вычисления оптимальной функции $\varphi(n_j)$. Выражение (5) в условно-непрерывном виде может быть записано следующим образом:

$$F_{1.H} = \int_0^{n_m} n \cdot \log_2 \varphi(n) dn + \lambda \int_0^{n_m} \varphi(n) dn \quad (7)$$

при $\int_0^{n_m} \varphi(n) dn = C, \quad (8)$

где C = const.

Согласно уравнению Эйлера [11], оптимальная функция $\varphi(n)_{opt}$, приводящая функционал (7) к его экстремальному значению, должна удовлетворить условию

$$F_2 = \frac{d\left\{n \cdot \log_2 \varphi(n) + \lambda \cdot \varphi(n)\right\}}{d\varphi(n)} = 0 . \quad (9)$$

С учетом выражений (7) и (8) получаем

$$\frac{n}{\varphi(n)} + \lambda = 0 . \quad (10)$$

Из выражения (10) находим

$$\varphi(n) = -\frac{n}{\lambda} . \quad (11)$$

С учетом выражений (8) и (11) получаем

$$-\int_0^{n_m} \frac{n}{\lambda} dn = C . \quad (12)$$

Из выражения (12) получим

$$\lambda = -\frac{n_m^2}{2C} . \quad (13)$$

С учетом выражений (10) и (13) получаем

$$\varphi(n)_{opt} = \frac{2C \cdot n}{n_m^2} . \quad (14)$$

Таким образом, при оптимальной функции $\varphi(n)_{opt}$, определяемой выражением (14), информационное содержание идеально кластерированного множества X, определяемое выражением (7), достигает экстремума. При этом экстремум является максимумом, так как выражение

$$F_2 = \frac{d^2\left\{n \cdot \log_2 \varphi(n) + \lambda \cdot \varphi(n)\right\}}{d\varphi(n)^2} \quad (15)$$

имеет отрицательное значение.

Сокращение признакового пространства

Как видно из выражения (14), оптимальная функция $\varphi(n)_{opt}$ зависит от переменного n и от параметра C, исходно задаваемого при решении задачи оптимальной кластеризации. Рассмотрим возможность сокращения признакового пространства в предложенном методе кластеризации. Предлагается следующий алгоритм сокращения признакового пространства.

1. Определяется максимальная величина функционала:

$$F_{1.H} = \int_{1}^{n_m} n \cdot \log_2 \varphi(n) dn . \quad (16)$$

С учетом выражений (14) и (16) имеем

$$F_{1.H.1.\max} = \int_{1}^{n_m} n \cdot \log_2 \frac{2C \cdot n}{n_m^2} dn . \quad (17)$$

2. Определяется реальная величина функционала (16).

С учетом выражения (16) и $\varphi=\varphi(n)_{real}$ получаем

$$F_{1.H.1.real} = \int_{0}^{n_m} n \cdot \log_2 \varphi(n)_{real} \, dn . \quad (18)$$

3. Критерий сокращения признакового пространства имеет вид $\beta = \frac{F_{1.H.1.\max} - F_{1.H.1.real}}{F_{1.H.1.\max}}$.

Принимается, что по достижении условия $\beta \geq a_0$, где a_0 – заранее заданное число, $a_0 \leq 1$, элементы x_i с количеством состояний, определяющих величину $F_{1.H.1.real}$, могут быть исключены из рассмотрения.

Заключение

Таким образом, предлагаемый метод оптимальной энтропийной кластеризации, базируясь на известном принципе «малая величина энтропии соответствует большому количеству информационного содержания», позволяет формировать режим не только оптимальной кластеризации, но и сокращения признакового пространства. Предлагаемый алгоритм может быть реализован в среде MATLAB методом последовательных приближений. Преимуществом среды MATLAB является возможность быстрой по сравнению с ФОРТРАН разработки рабочего алгоритма, а также альтернативных решений с использованием существующего пакета программ в этой среде.

Таким образом, были предложены метод оптимальной энтропийной кластеризации высокоразмерных данных в информационных системах с его математическим обоснованием, а также методика сокращения признакового пространства высокоразмерных данных при первичной обработке.

Литература

1. Бабичев С.А. Оптимизация процесса предобработки информации в системах кластеризации высокоразмерных данных // Radioelektronika, informatika, управління. 2014. № 2. С. 135–142.
2. Li H., Zhang K., Jiang T. Minimum entropy clustering and applications to gene expression analysis. URL: https://www.ncbi.nlm.nih.gov/pubmed/16448008 (дата обращения: 30.03.2017).
3. Santos J.M., Sa J.M., Alexandre L.A. LEGClust – a clustering algorithm based on layered entropic subgraphs. IEEE Transactions on pattern analysis and machine intelligence, 2008, vol. 30, no. 1, pp. 1–13.
4. Smieja M., Nakoneczny S., Tabor J. Fast entropy clustering in sparse high dimensional binary data. URL: http://ww2.ii.uj.edu.pl/~smieja/publications/SEC.pdf (дата обращения: 30.03.2017).
5. Singh B., Kushwaha N., Vyas O.P. A feature subset selection technique for high dimensional data using symmetric uncertainty. Jour. of Data Analysis and Information Processing, 2014, no. 2, pp. 95–105.
6. Выбор компонентов (интеллектуальный анализ данных). URL: http://msdn.microsoft.com/ru-ru/library/ms175382(d=printer).aspx (дата обращения: 30.03.2017).
7. Zimmermann A. Objectively evaluating interestingness measures for frequent item set mining. URL: http://zimmermanna.users.greyc.fr/papers/international-workshops/pakdd2013objectively-evaluating.pdf (дата обращения: 30.03.2017).
8. Sahar S. What is interesting: studies on interestingness in knowledge discovery. URL: http://www.cs.tau.acil/~mansour/students/SigalSaharPhD.pdf (дата обращения: 30.03.2017).
9. Tew C., Giraud-Carrier C., Tanner K., Burton S. Behavior – based clustering and analysis of interestingness measures for association rule mining. URL: http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Slides/10.1--7_s10618-013-0326-x.pdf (дата обращения: 30.03.2017).
10. Malik H.H., Kender J.R. Instance Driven Hierarchical Clustering of Document Collections. URL: http://www.ke.tu-darmstadt.de/events/LeGo-08/8.pdf (дата обращения: 30.03.2017).
11. Эльгольц Л.П. Дифференциальные уравнения и вариационное исчисление. М.: Наука, 1974. 432 с.

Software & Systems
DOI: 10.15827/0236-235X.120.643-646

Received 31.03.17
2017, vol. 30, no. 4, pp. 643–646

OPTIMUM ENTROPY CLUSTERING IN INFORMATION SYSTEMS

B.G. Askerova [1], *Ph.D. (Engineering), Associate Professor, Bahar287@mail.ru*

[1] *Azerbaijan State University of Oil and Industry, Azadlyg Ave. 20, Baku, AZ1010, Azerbaijan*

Abstract. The paper researches the possibility of developing a new method for data clustering in information systems. Clustering is a process of searching possible groups in a given set using signs of similarity or difference between elements of this set. The existing entropy clustering method includes an information theoretic approach to a clustering task. The paper suggests a clustering method based on an entropy approach to selecting message items.

The paper suggests a method of optimum entropy clustering of high-dimensional data in information systems. It also gives mathematical grounding of the method.

The suggested method of optimum entropy clustering is based on the known principle “low entropy corresponds to big information content”. This make it possible to form an optimum clustering regime, as well as an attribute space reduction regime.

The paper proposes a method for calculating a level of clustering optimality. It also describes a method for reducting attribute space of high-dimensional data upon their initial processing.

Keywords: clustering, optimization, high-dimensional data, information systems, entropy.

References

1. Babichev S.A. Optimization of information preprocessing in clustering systems of high dimension data. *Radio Electronics, Computer Science, Control*. 2014, no. 2, pp. 135–142 (in Ukr.).

2. Li H., Zhang K., Jiang T. *Minimum entropy clustering and applications to gene expression analysis*. Proc. Conf. Computational Systems Bioinformatics (CSB 2004). 2004, IEEE, pp. 142–151.

3. Santos J.M., Sa J.M., Alexandre L.A. LEGClust – A clustering algorithm based on layered entropic subgraphs. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. 2008, vol. 30, no. 1, pp. 1–13.

4. Smieja M., Nakoneczny S., Tabor J. *Fast entropy clustering in sparse high dimensional binary data*. Available at: http://ww2.ii.uj.edu.pl/~smieja/publications/SEC.pdf (accessed March 30, 2017).

5. Singh B., Kushwaha N., Vyas O.P. A feature subset selection technique for high dimensional data using symmetric uncertainty. *Jour. of Data Analysis and Information Processing*. 2014, no. 2, pp. 95–105.

6. *Vybor komponentov (intellektualny analiz dannykh)* [Selecting Components (Data Mining)]. Available at: http://msdn.microsoft.com/ru-ru/library/ms175382(d=printer).aspx (accessed March 30, 2017).

7. Zimmermann A. Objectively evaluating interestingness measures for frequent item set mining. *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining*, 2013, pp. 354–366.

8. Sahar S. What is interesting: studies on interestingness in knowledge discovery. *PhD Thesis*, Tel-Aviv Univ., 181 p.

9. Tew C., Giraud-Carrier C., Tanner K., Burton S. Behavior – based clustering and analysis of interestingness measures for association rule mining. *Jour. of Data Min. Knowl. Disc*., 2014, vol. 28, no. 4, pp. 1004–1045.

10. Malik H.H., Kender J.R. Instance driven hierarchical clustering of document collections. *Proc. Conf. Local Patterns to Global Models (ECML/PKDD-08)*, 2008.

11. Eltsgolts L.P. *Differential Equations and Variational Calculus*. Moscow, Nauka Publ., 1974, 432 p. (in Russ.).

Примеры библиографического описания статьи

1. Аскерова Б.Г. Оптимальная энтропийная кластеризация в информационных системах // Программные продукты и системы. 2017. Т. 30. № 4. С. 643–646. DOI: 10.15827/0236-235X.120.643-646.

2. Askerova B.G. Optimum entropy clustering in information systems. *Programmnye produkty i sistemy* [Software & Systems]. 2017, vol. 30, no. 4, pp. 643–646 (in Russ.). DOI: 10.15827/0236-235X.120.643-646.