Machine Learning and Artificial Intelligence
J.-L. Kim (Ed.)

doi:10.3233/FAIA230768

Research on a Symbolic Cluster Analysis Algorithm Approaching to the Optimal Value

Fang ZANG[1]
Hunan Urban Vocational College, Changsha, 410137, China

Abstract. Nowadays, there is a large amount of symbolic data in various fields, and people can fully utilize these symbolic data for clustering, providing a better foundation and direction for data mining and analysis. Currently, clustering algorithms for symbol data have emerged one after another, but there are still shortcomings in computational cost and algorithm robustness. Therefore, it is urgent to study an algorithm with stable clustering results, less time consumption, and low I/O overhead. The following proposes a symbolic cluster analysis algorithm that approaches to the optimal value. It reduces the size of the original data by generating a symbolic association graph from a large number of symbolic data samples, effectively solves the problem of high computing costs caused by the huge amount of data, and proves the clustering effect of the algorithm through empirical analysis.

Keywords. Symbolic data, Symbolic association graph, Clustering analysis

1. Introduction

In recent years, cluster analysis [1] has been widely applied in fields such as statistics, image processing, healthcare, information search, biotechnology, machine learning, etc [2]. The role of cluster analysis is to discover data objects and their relationships based on data, and group these data to make the affinity between data within the group higher than that between other data. The higher the affinity within a group and the greater the distance between groups, the better the clustering effect.

Data objects can be roughly divided into numerical data, symbolic data, mixed data and other types based on their properties [3]-[4]. Among them, symbol data is usually aggregated from large datasets, used to hide specific details of items and convert large amounts of data into analyzable quantities. Symbolic data is different from numerical data. Numerical data can describe the affinity between data using mathematical methods such as Euclidean distance. However, there is no stable mathematical relationship between symbolic data, and distance formulas cannot be used to measure the affinity between the two, resulting in the inability to use general numerical clustering methods to process symbolic data. In today's information data explosion, there is a large amount of symbolic data everywhere, so the research on symbolic data has become of great significance.

[1] Corresponding Author, Fang Zang, Hunan Urban Vocational College, Xuhui Cang Jun, Yuelu District, Changsha, Hunan Province, China; E-mail: 1174899@qq.com.

2. Research progress in symbolic data clustering algorithms

So far, there has been an endless stream of research on clustering algorithms for symbol data, with general directions based on information content, probability statistics, and different measures.

The principle of information based symbolic clustering algorithm [5] is to use information entropy to express the uncertainty in data description, and then calculate the consistency of clustering partition on data description in feature space and symbol space, and achieve clustering partition of data set under the iterative computing framework, so that the clustering partition results are more accurate and robust. As the name implies, the clustering algorithm based on probability and statistics uses the principle of probability theory to deal with clustering problems, usually using Bayesian rule and maximum likelihood estimation in probability and statistics. For example, there are n types of data in the original dataset with a certain number of probability distributions, and these data are classified into different classes with a certain probability, and the size of the probability value determines the division into different groups. The clustering algorithm based on dissimilarity measure [6] is research on the distance between symbol data, mainly studying the distance between data and data, as well as the distance between data and different classes [7].

In 1987, the COBWEB algorithm [8], also known as the simple incremental concept clustering algorithm, was introduced. It uses a tree graph to describe hierarchical classification, where a node is a concept and property value is used to describe input objects [9]. In 1995, Ralambondrainy proposed a method of converting symbol attributes into binary attributes and clustering them based on the K-Means algorithm [10]. In 1997, a function of affinity and alienation based on location, span and content was proposed. The operation of this algorithm largely depends on the hierarchical clustering method of prior knowledge in specific fields [11]. In 1998, Huang first proposed the famous K-Modes algorithm for symbolic cluster analysis [12], which uses simple dissimilarity matching metrics. However, this algorithm has problems such as large initialization errors for k-clusters and inaccurate classification of large amounts of data samples by k-nearest neighbors. In 2002, the distance based discrete cluster analysis COOLCAT incremental algorithm [13] is mainly aimed at minimizing the amount of clustering information, that is, when the initial clustering symbol data set is specified, as long as its total clustering information is the smallest, the algorithm locates the next mode of clustering. In 2005, the ROCK algorithm for measuring the similarity between two symbol data points was proposed by the Sudipto team. The core idea of the ROCK algorithm [14] is based on the affinity measurement of "links". When we consider whether to merge cluster X and cluster Y, we calculate the number of links between the data points in the two clusters. In 2016, Sharma et al. proposed the concept of General Similarity (GSM) based on the ROCK algorithm [15]. The principle is to convert multiple general metrics into individual formulas with parameters, and then use the ROCK algorithm to verify which one is the most effective. In the same year, based on the research on Shannon entropy clustering algorithms, Professor Sharma proposed the TEC algorithm, which performs better by describing all attributes as power-law algorithms [15]. Based on the research of typical K-means algorithm and K-Medoids algorithm, Nguyen proposed an algorithm in 2016 that is an extension of the k-means algorithm and can automatically measure the impact of various attributes on clustering [16]. In 2017, Ding et al. [17] proposed that in order to weaken the interference of data input sequence on the clustering effect, the random sequence of the original data and the

attribute weight were used to reshape the clustering algorithm, and on the basis of attribute entropy, they proposed a calculation method to distinguish the attribute weight. In 2019, Jia et al. [18] improved the K-Medoids algorithm by ignoring the differences between attributes and being greatly affected by the original center points, and proposed a pre clustering based original center selection method. In 2019, Wang et al. [19] improved on the above algorithms and proposed a k-modes KNN algorithm. In 2019, Muhammadu's team proposed an algorithm to solve symbolic cluster analysis with evidence clustering, namely ECM algorithm [20]. In this algorithm, a dissimilarity measure was defined, and the iterative optimization algorithm was used as the basis for clustering grouping. The above main symbolic cluster analysis algorithms are summarized as Table 1:

Table 1. Comparison of main symbol cluster analysis algorithms.

Algorithm name	Algorithm features	Classify
COOLCAT	When a symbol dataset with initial clustering is specified, as long as the expected entropy of its total clustering is the smallest, the algorithm locates the next mode of the clustering.	Based on information content
TEC	When all attributes are described as power-law, it outperforms clustering algorithms based on Shannon entropy.	
Ng' s K-Modes	An algorithm that automatically measures the impact of various attributes on clustering.	
COBWEB	Hierarchical clustering is created in the form of a classification tree, and its input objects are described by classification attribute value pairs.	Based on probability statistics
ECM	Define a dissimilarity measure, successfully introduce alternating minimization to obtain partitions, and apply evidence clustering to symbolic cluster analysis for the first time.	
ROCK	Based on the affinity measure of "link", when we consider whether to merge cluster X and cluster Y, we calculate the number of links between two data points in the two clusters.	Based on dissimilarity measure

At present, the symbolic cluster analysis algorithm has achieved good results, but there are still some common shortcomings. In summary, it can be summarized as follows: the scalability is not strong, and the clustering effect is not ideal when the data set increases. The main manifestations are long computation time and high computation cost. Some algorithms solve the problem of long computation time, but the size of parameters leads to significant differences in clustering results. In order to improve the efficiency and robustness of the symbolic cluster analysis algorithm, a symbolic clustering algorithm approaching to the optimal value is proposed. The core idea of the algorithm is to use the symbolic association graph on the original data set to reduce the size of the data set as much as possible, so as to reduce the I/O overhead and improve the operation efficiency and clustering effect.

3. A symbolic clustering algorithm for approaching optimal values

A symbolic clustering algorithm that uses correlation graphs to replace a large number of symbolic data samples and approach the optimal clustering results in order to reduce the size of the dataset. The key steps are as follows:

1) Determine the correlation diagram based on the original data [21][22][23].

2) Grouping association graphs based on certain graph splitting algorithms.

3) Based on different symbols, the grouping of datasets is determined by finding the maximum probability between them as the basis for class grouping.

Assuming the original symbol dataset is as Table 2:

Table 2. Original Data Set.

Student name	Sex	Province
Li Xiaoying	male	Hunan
Zhang Daixuan	female	Hunan
Wang Luyuan	male	Hubei

Below is a legend to illustrate the main steps of the algorithm:

1) Establish an association diagram between symbol data, as shown in Figure 1.

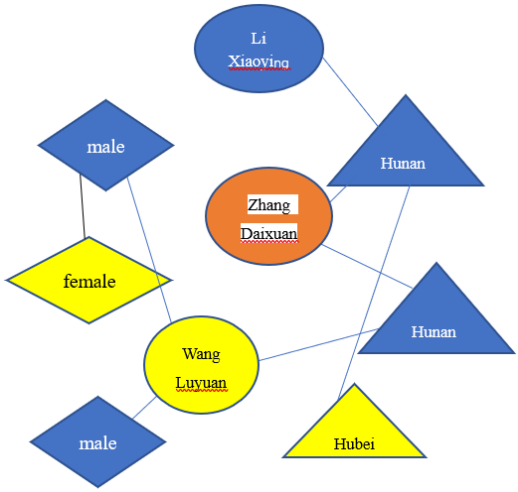

Figure 1. Correlation diagram between symbol data.

2) Using graph splitting algorithm to segment association graphs.

3) Using the highest probability as the clustering basis to obtain the clustering results, as shown in Figure 2.

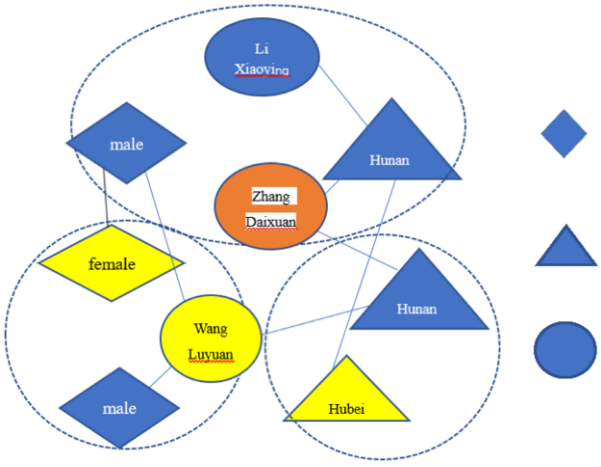

Figure 2. Schematic diagram of association graph segmentation & Clustering Results.

Explanation: Different attributes are presented in different shapes, the same value of the same attribute is represented in the same color, and different values are identified in different colors, such as blue for males and yellow for females. Hunan Province is marked in blue, while Hubei Province is marked in yellow.

3.1. Establishment of symbolic association diagram

The purpose of establishing a symbolic association graph is to decompose a large-scale dataset into several small-scale datasets, in order to shorten algorithm time, reduce computational costs, and improve operational efficiency.

Table 3 shows some symbol attribute information for student information. The sample dataset $U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7\}$, consists of 7 objects. $A = \{a_1, a_2, a_3, a_4, a_5\}$ represents the 5 attributes in Table 3, namely Score, Classroom performance, looks, height, and class. Then use $u_i = \{u_{i1}, u_{i2}, \cdots u_{im}\}$represents the i-th object composed of m attributes, and for this purpose, matrix rows and columns are used to represent the symbol object. $D(a_j) = \{a_{j1}, a_{j2}, \cdots a_{jn_j}\}$represents the range of values for the j-th attribute. For example, the score attribute has values of excel, good, or pass, indicating that the number of attribute values is 3. Use n_j represents that the number of Classroom performance attribute values is also 3, the number of looks attribute values is 5, the number of height attribute values is 3, and the number of class attribute values is 2. Class represents the class to which the data object belongs. Graphic express refers to the graphical representation in a symbolic relationship diagram, where different graphics represent different attributes, with the same color representing the same attribute value.

Table 3. Example of Symbol Attributes in Student Information System.

Obj.	Score	GP.	Classroom performance	GP.	looks	GP.	height	GP.	class
X_1	good		poor grades		pretty		medium		B
X_2	good		preferably		charming		short		A
X_3	good		poor grades		preferably		tall		B
X_4	pass		preferably		pretty		medium		A
X_5	Excellent		outstanding		ordinary		tall		A
X_6	pass		outstanding		deformity		tall		A
X_7	Excellent		outstanding		ordinary		tall		A

The above symbolic data in Table 3 cannot use the distance formula of numerical data to calculate its affinity, and use this as the basis for clustering. Because the traditional symbolic cluster analysis algorithm has certain defects in computing cost, algorithm robustness, and scalability, therefore, the traditional symbolic cluster analysis algorithm cannot be used, but the form of association graph can be used to greatly reduce the size of the data set. The symbolic data in Table 3 above is represented by matrix representation as:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 1 & 3 & 3 & 1 \\ 2 & 2 & 1 & 1 & 2 \\ 3 & 3 & 4 & 3 & 2 \\ 2 & 3 & 5 & 3 & 2 \\ 3 & 3 & 4 & 3 & 2 \end{bmatrix}$$

The columns in the matrix correspond to the attribute values in Table 2, with 1, 2, 3,… representing different symbol data values in each attribute column. The meaning of 1 in the first attribute column is different from that of 1 in the second attribute column. Similarity matrix $B \in R^{v\times v}$ Define the similarity matrix formula as:

$$B_{a_{ij},a_{kl}} = \frac{|\{u_{ij}=a_{ij}\}\cap\{u_{ik}=a_{kl}\}|}{\sqrt{|\{u_{ij}=a_{ij}\}||\{u_{ik}=a_{kl}\}|}} \tag{1}$$

Explanation: $v = \sum_{i=1}^{m} n_{j,}$means which is the sum of the number of values for each attribute.i、 $k \in \{1,\ 2,\ 3, \ldots, m\}$attribute set, j、 $l \in \{1,\ 2,\ 3, \ldots, v\}$ attribute value set. $\{u_{ij} = a_{ij}\}$means that the attribute value is the set of a_{ij}, $\{u_{ik} = a_{kl}\}$ means that the attribute value is the set of a_{kl}. In Table 2, there are 16 attribute values of $v = (3 + 3 + 5 + 3 + 2) = 16$, so the similarity matrix representation is $B \in R^{16\times16}$. According to Eq. (1), it is obtained that:

$$B_{1,1} = \frac{|\{u_{11,}u_{21,}u_{31}\}\cap\{u_{11,}u_{21,}u_{31}\}|}{\sqrt{|\{u_{11,}u_{21,}u_{31}\}|\times|\{u_{11,}u_{21,}u_{31}\}|}} = \frac{3}{\sqrt{3\times3}}=1 \tag{2}$$

Explanation: $B_{1,1}$ is represented as the value of the first row and first column of the similarity matrix, and the numerator in the result is the intersection of the set of the first attribute value in all attribute 1 and the set of the first attribute value in all attribute 1. Since there are three attribute values in the two sets, the numerator is 3.

$$B_{1,2} = \frac{|\{u_{11,}u_{21,}u_{31}\}\cap\{u_{41,}u_{61}\}|}{\sqrt{|\{u_{11,}u_{21,}u_{31}\}|\times|\{u_{41,}u_{61}\}|}} = \frac{0}{\sqrt{3\times2}}=0 \tag{3}$$

Explanation: $B_{1,2}$ is represented as the values in the first row and second column of the similarity matrix, and the numerator in the result is the intersection of the first attribute value set in all attribute 1 and the second attribute value set in all attribute 1. Since the attribute values in the two sets are different, the intersection is 0.

$$B_{2,2} = \frac{|\{u_{41,}u_{61,}\}\cap\{u_{41,}u_{61}\}|}{\sqrt{|\{u_{41,}u_{61,}\}|\times|\{u_{41,}u_{61}\}|}} = \frac{2}{\sqrt{2\times2}}=1 \tag{4}$$

Explanation: $B_{2,2}$is represented as the value in the second row and second column of the similarity matrix, and the numerator in the result is the intersection of the second attribute value set in all attribute 1 and the second attribute value set in all attribute 1. Since the attribute values in the two sets are the same and there are two, the intersection is 2. By analogy, we obtain $B \in R^{16\times16}$ All values in, and the matrix is symmetric.

3.2. Symbolic relationship graph clustering for approaching optimal values

Using MATLAB software in matrix calculation, symbol data is divided into several categories using commonly used clustering algorithms. $K = \{k_1, k_2, k_3, \ldots, k_n\}$, k_n is the nth set among all symbol sets. So the class label (CL) of each symbol value is represented as:$CL(\mathrm{a_{il}}) = n, if\ \mathrm{a_{il}} \in k_n, i \in [1, m], l \in [1, n_j]$.

The above correlation graph can better present the affinity between symbol attribute values, and combine traditional clustering algorithms to obtain class labels, thereby obtaining a clustering result that approaches the optimal value. Define and label the class label in (1) again, and after labeling, it represents the following: $u_{iq} = CL(\mathrm{a_{il}}), if\ u_{iq} = \mathrm{a_{il}}, i \in [1, n], q \in [1, m]$.

Then find the most probability of each row in the label matrix as the basis for class division, which directly affects the final symbolic cluster analysis results. Using arg max to describe the clustering label L as:$L(u_i) = \arg\max |\{u_{iq} = k, q \in [1, m]|$.

Use $P = (T, S)$ to represent an undirected weighted association graph, T represents the set of all symbolic attribute values, and S represents the affinity weighting between any two attribute values. Like $S_{a_{ij,}a_{kl}}$ represents the affinity weighting between the j_{th} attribute value of the i_{th} attribute column and the l_{th} attribute value of the k_{th} attribute column. Based on the radial basis function $k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$ Formula, based on the characteristics of symbol attribute values, defines the weighted affinity values between symbol relationship graph attributes as:

$$S_{a_{ij,}a_{kl}} = e^{\left(A+\frac{s_{a_{ij}} \cdot s_{a_{kl}}}{\sigma\sqrt{\left|s_{a_{il}}\right|\left|s_{a_{kl}}\right|}}\right)}, \mathrm{A}=-\frac{1}{\sigma} \tag{5}$$

Explanation: Based on the properties of the row column in the matrix, $s_{a_{il}}$ in the formula a_{il} represents a in the similarity matrix The line where il is located, σ as a Gaussian kernel parameter (set to 0.0897 in this experiment), the higher the S value, the higher the similarity of symbol attribute values. Conversely, the larger the S weight, the lower the similarity of symbol attribute values.

According to Eq. (5), obtain the symbol association diagram, as shown in Figure 3:

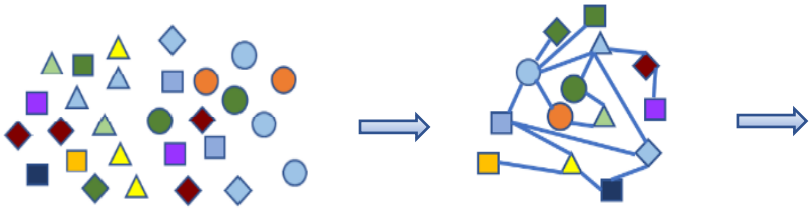

Figure 3. Formation process of symbol association diagram.

Finally, perform graph splitting. The basis for graph splitting in the algorithm is the weight of the edges in the correlation graph, with the weight $S_{a_{ij,}a_{kl}}$The larger, the lower the similarity and the easier it is to split, resulting in high similarity being divided into the same subgraph and low similarity being split into different subgraphs (Figure 4).

Figure 4. Splitting diagram of symbol association diagram.

3.3. Algorithm analysis

Time cost analysis of this algorithm. Due to the symmetry of the similarity matrix composed of symbol elements, the time consumption for obtaining the similarity matrix is $O(v^2/2)$, and the time consumption for obtaining the relationship graph is the same as before. The time consumption for obtaining the class label is $O(n)$, and the time consumption for obtaining the final clustering result is $O(m)$. Therefore, the total time consumption of this algorithm is $O(v \times v + n + m)$ .The entire algorithm process is shown in Figure 5. Input: Dataset and number of clusters parameter n. Output: Cluster label L.

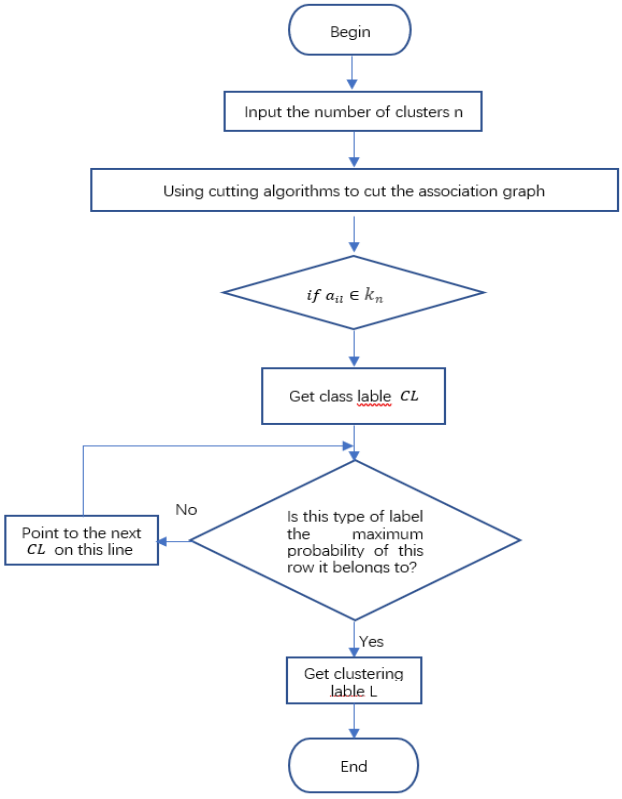

Figure 5. Algorithm flowchart for approaching the optimal value.

4. Experimental instructions

4.1. Experimental preparation and evaluation indicators

The experiment was conducted in a 3.4GHz CPU, 16GB Memory, Win12OS environment, and MATLAB 2020 b environment. The purpose of this experiment is to verify that the clustering performance of the algorithm proposed in the article approaches optimal, mainly comparing the quality of clustering and the computational time required for clustering. NMI (Normalized Mutual Information), ARI (Adjusted Rand Index) and PE (precision) are used for evaluation. Multiple datasets of different scales were selected from the machine learning library, as shown in the Table 4:

Table 4. Experimental Dataset.

Data Sets	Instance	Features	Classes
Pepper	47	35	4
Aquarium	101	16	7
LSC	148	18	8
surgery	366	33	6
Floor-vote	435	16	2
lung cancer lung cancer	699	10	2
DeoxyriboNucleic Acid	3190	60	3
insect	8124	22	2
tc_rich_KT	2310	100	7
tc_rich_KFC	3780	100	7
tc_rich_NOMSI	5001	100	10
tc_rich_KTUSC	7790	100	26
tc_rich_EW	20010	100	26

The evaluation of clustering quality mainly measures the similarity between the actual class labels in the dataset and the clustering results. As provided above, a dataset with N (13) data objects is provided. Specify $K = \{k_1, k_2, k_3, \dots, k_n\}$ to indicate that there are n clustering results, $n_{kl} = |k_n \cap L_l|$ represents the k_n group in the clustering results and actual class label the L_l group how many common symbol attribute values are there. $b_k = \sum_{l=1}^{N} n_{kl}$, $d_l = \sum_{k=1}^{N} n_{kl}$.

The Normalized Mutual Information [24] evaluation index is expressed as:

$$NMI = \frac{2\sum_k \sum_l n_{kl} log \frac{n_{kl} N}{b_k d_l}}{-\sum_k b_k log \frac{b_k}{N} - \sum_l d_l log \frac{d_l}{N}} \tag{6}$$

The evaluation index for Adjusted Rand Index [25] is expressed as:

$$ARI = \frac{\sum_{kl}\binom{n_{kl}}{2} - \left[\sum_k \binom{b_k}{2} \sum_l \binom{d_l}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_k \binom{b_k}{2} + \sum_l \binom{d_l}{2}\right] - \left[\sum_k \binom{b_k}{2} \sum_l \binom{d_l}{2}\right] / \binom{N}{2}} \tag{7}$$

The evaluation index for accuracy [26] is expressed as:

$$PE = \frac{1}{n} \sum_{k=1}^{n} \frac{max_{l=1\dots n'} n_{kl}}{b_k} \tag{8}$$

4.2. Analysis of experimental results

The following Table 5-Table 7 shows the data of various main symbol data clustering algorithms and symbol data clustering algorithms based on symbol association graphs mentioned earlier, which measure clustering effectiveness through three evaluation indicators: NMI, ARI, and PE.

Table 5 shows that the symbol data clustering algorithm based on symbol association graph has the best clustering performance in 8 out of 13 data objects in NMI evaluation, and the clustering performance of the other 5 is best dispersed among other algorithms. Although the clustering effect is not ideal on DeoxyriboNuclear Acid and LSC objects, it cannot be denied that the symbol data clustering algorithm based on symbol association graph is an algorithm that approaches the optimal value as a whole.

Table 5. Standardized Mutual information NMI evaluation of clustering results of different algorithms.

Datasets	COOL CAT	ROCK	Ng's K-Modes	San's K-Modes	ACE	This algorithm
Lung Cancer	0.3801	0.8063	0.7129	0.5909	0.6589	0.8355
Pepper	0.7589	0.5283	0.3390	0.7589	0.7589	0.6853
Aquarium	0.1487	0.5331	0.1487	0.1487	0.4879	0.5523
Floor-vote	0.1788	0.9780	0.1788	0.1788	0.1788	0.6637
Surgery	0.0657	0.3814	0.1657	0.1657	0.1657	0.4770
DeoxyriboNucleic Acid	0.4755	0.2545	0.0530	0.1295	0.1921	0.0384
Insect	0.1701	0.1021	0.1701	0.667	0.4064	0.3119
LSC	0.1638	0.6987	0.1638	0.438	0.121	0.174
tc_rich_KT	0.466	0.4936	0.466	0.274	0.2807	0.5136
tc_rich_EW	0.2826	0.3381	0.2826	0.5037	0.143	0.6731
tc_rich_KFC	0.5901	0.3406	0.6741	0.2826	0.4844	0.6831
tc_rich_NOMSI	0.741	0.3882	0.7901	0.502	0.5326	0.8954
tc_rich_KTUSC	0.5648	0.4079	0.5648	0.2484	0.4406	0.6032

Table 6 shows that the ARI evaluation of the symbol data clustering algorithm based on symbol association graph has the best clustering performance in 8 out of 13 data objects, while the ROCK algorithm has the best clustering performance in the other 5 objects. Although the clustering effect on LSC objects is not ideal, it cannot be denied that the symbol data clustering algorithm based on symbol association graph is an algorithm that approaches the optimal value as a whole.

Table 6. ARI evaluation of adjusted Rand coefficients for clustering results of different algorithms.

Datasets	COOLCAT	ROCK	Ng' s K-Modes	San's K-Modes	ACE	This algorithm
Lung Cancer	0.4513	0.6710	0.7240	0.6240	0.5040	0.7335
Pepper	0.1227	0.4325	0.6529	0.7227	0.6227	0.5542
Aquarium	0.3833	0.4447	0.5833	0.6823	0.5733	0.8533
Floor-vote	0.1244	0.9908	0.3454	0.6204	0.7244	0.5487
Surgery	0.0270	0.2545	0.2709	0.2709	0.2709	0.4076
DeoxyriboNucleic Acid	0.3712	0.0474	0.3421	0.5326	0.0437	0.5730
Insect	0.0088.	0.6430	0. 0088	0.0088	0.0562	0.6241
LSC	0.1381	0.4195	0.1381	0.1381	0.0638	0.1002
tc_rich_KT	0.2739	0.1021	0.2739	0.5739	0.4323	0.6051
tc_rich_EW	0.1486	0.3908	0.1486	0.1486	0.4096	0.4198
tc_rich_KFC	0.0747	0.1023	0.0747	0.0747	0.3722	0.2369
tc_rich_NOMSI	0.2124	0.5637	0.2246	0.2124	0.6072	0.6534
tc_rich_KTUSC	0.3321	0.3744	0.2453	0.2475	0.4539	0.4627

Table 7 shows that the ARI evaluation of the symbol data clustering algorithm based on symbol association graph has the best clustering performance in 9 out of 13 data objects, and the clustering performance of the other 4 objects is mainly concentrated on the ROCK algorithm.

Table 7. Precision PE evaluation of clustering results of different algorithms.

Datasets	COOLCAT	ROCK	Ng' s K-Modes	San's K-Modes	ACE	This algorithm
Lung Cancer	0.8187	0.8933	0.6499	0.6499	0.9471	0.9572
Pepper	0.9062	0.6258	0.3616	0.3616	0.7810	0.8240
Aquarium	0.7064	0.6412	0.4058	0.4094	0.6852	0.7107
Floor-vote	0.7594	0.8075	0.6137	0.6137	0.8799	0.9110
Surgery	0.7518	0.548	0.3059	0.3059	0.7647	0.8273
DeoxyriboNucleic Acid	0.6899	0.6272	0.3344	0.5179	0.4247	0.6972
Insect	0.6961	0.4395	0.5179	0.3863	0.4527	0.7820
LSC	0.4329	0.7913	0.3851	0.4608	0.6463	0.6581
tc_rich_KT	0.6541	0.6684	0.4291	0.4070	0.4167	0.7759
tc_rich_EW	0.6578	0.9264	0.0407	0.1300	0.2089	0.6200
tc_rich_KFC	0.3185	0.5747	0.1429	0.3871	0.1796	0.3365
tc_rich_NOMSI	0.6338	0.5780	0.1000	0.3987	0.1467	0.7675
tc_rich_KTUSC	0.6584	0.5343	0.3850	0.3630	0.5975	0.6708

Table 8 shows the comparison of clustering times among various algorithms. Although the clustering time on the larger object is slightly inferior to the Ng's K-Modes algorithm, it can still be considered that this algorithm has a significant advantage in clustering time.

Table 8. Comparison of time consumption of various clustering algorithms.

Datasets	COOLCAT	ROCK	Ng's K-Modes	San's K-Modes	ACE	This algorithm
lung cancer	0.655	131.001	0.624	0.720	905.941	0.540
Pepper	0.61	0.296	0.880	0.45	0.49	0.233
Aquarium	0.31	1.186	0.25	0.30	3.093	0.131
Floor-vote	0.366	36.328	0.309	0.779	211.132	0.208
surgery	2.525	23.218	0.401	2.081	847.303	1.288
DeoxyriboNucleic Acid	43.138	3405	58.98	64.141	1587.8	8.370
insect	37.942	625.8	1.921	1.468	503.758	1.052
LSC	0.769	2.514	0.769	0.459	11.124	0.228
tc_rich_KT	12.703	87251	10.299	56.430	555.79	9.143
tc_rich_EW	2113.3	10763	19.21	79.372	7803	12.4
tc_rich_KFC	39.397	9811	70.855	69.18	357.66	19.5
tc_rich_NOMSI	84.002	4536	33.75	43.27	307.04	24.2
tc_rich_KTUSC	250.995	1052	18.658	25.609	964.8	9.3

The data shown in the following two tables is a comparison of the clustering performance of traditional clustering algorithms based on symbolic association graphs. The now column shows the clustering performance of the data after using symbolic association graphs, while the before column shows the clustering performance of the data that has not been used. If the clustering performance of the two is better compared, a shadow is used to indicate. As shown in Table 9, the horizontal data of lung cancer, Floor vote, LSC, tc_ rich_ KT, the clustering effect of tc_ rich_ KTUSC has significantly improved. The longitudinal data shows that the three traditional clustering algorithms using symbol association graph-based methods have overall better clustering performance than the previous ones.

Table 9. Standardized Mutual information NMI evaluation of clustering results after using symbolic association graph in traditional clustering algorithm.

Datasets	SPC		K-Modes		HC	
	now	before	now	before	now	before
Lung Cancer	0.8356	0.7335	0.8742	0.6833	0.763	0.52
Pepper	0.3054	0.4468	0.3054	0.6553	0.1228	0.6738
Aquarium	0.5024	0.4158	0.2694	0.6707	0.3834	0.4381
Floor-vote	0.6638	0.6138	0.5632	0.4905	0.3922	0.0037
Surgery	0.3197	0.1477	0.2063	0.4149	0.217	0.0045
DeoxyriboNucleic Acid	0.0385	0.5188	0.0341	0.0272	0.16	0.214
Insect	0.3119	0.6355	0.554	0.0604	0.0088	0.737
LSC	0.4257	0.0174	0.1228	0.0505	0.1381	0.027
tc_rich_KT	0.5286	0.4136	0.5104	0.4607	0.2624	0.252
tc_rich_EW	0.1759	0.1573	0.1823	0.1205	0.0056	0.19
tc_rich_KFC	0.1306	0.2839	0.1042	0.1024	0.893	0.451
tc_rich_NOMSI	0.4895	0.4522	0.3098	0.3972	0.452	0.171
tc_rich_KTUSC	0.42	0.2603	0.569	0.4709	0.3503	0.344

As shown in Table 10, the horizontal data Floor vote, tc_ Rich_ The clustering effect of NOMSI has significantly improved. The longitudinal data shows that the three traditional clustering algorithms using symbol association graph-based methods also perform better overall than the previous clustering algorithms.

Table 10. ARI evaluation of the adjusted Rand coefficient of clustering results after the use of symbolic association diagrams in traditional clustering algorithms.

Datasets	SPC		K-Modes		HC	
	now	before	now	before	now	before
Lung Cancer	0.7335	0.8389	0.783	0.5925	0.937	0.6515
Pepper	0.5543	0.4436	0.5543	0.8192	0.5319	0.8298
Aquarium	0.5534	0.4134	0.4874	0.7716	0.5842	0.6436
Floor-vote	0.6068	0.5487	0.445	0.4318	0.8138	0.6161
Surgery	0.4077	0.3264	0.354	0.5797	0.3579	0.3224
DeoxyriboNucleic Acid	0.0573	0.6357	0.1186	0.053	0.5191	0.5191
Insect	0.2411	0.6349	0.775	0.047	0.5643	0.5224
LSC	0.1002	0.4324	0.0941	0.2009	0.4257	0.4257
tc_rich_KT	0.6051	0.596	0.6415	0.6154	0.4489	0.4835
tc_rich_EW	0.4198	0.2943	0.3263	0.327	0.092	0.0747
tc_rich_KFC	0.2369	0.2841	0.1976	0.1849	0.1429	0.1444
tc_rich_NOMSI	0.6534	0.4559	0.6992	0.5354	0.198	0.1018
tc_rich_KTUSC	0.5089	0.4627	0.8014	0.7090	0.3424	0.1507

5. Conclusion

The above research aims to address the common shortcomings of existing symbol data clustering algorithms, such as long time consumption and unstable clustering results, and propose an improved symbol clustering algorithm that approaches the optimal value. The algorithm starts with the original data set, uses symbolic association graph to represent symbolic objects, finds out the affinity between symbols, and uses Spectral clustering, k-mean, Hierarchical clustering graph splitting algorithms to group the association graph. The above clustering algorithms were compared using clustering evaluation indicators, and the results showed that the optimal values of this algorithm on the experimental dataset were close, proving that this algorithm has advantages in solving problems such as weak processing power, high computational costs, and poor running speed for large amounts of data. However, there is still a large amount of complex symbol data in the

real world, and further research on high-dimensional symbol data association graphs is needed.

References

[1] Li JH. Research on parallelization of AIS data mining algorithm based on MapReduce Liaoning: Dalian Maritime University, 2022
[2] Liu R. Research and application of cluster analysis algorithm optimization in data mining Heilongjiang: Harbin Institute of Technology, 2004 DOI: 10.76666/d. Y656271
[3] Li X, Bai L. A clustering integration algorithm based on mixed data representation. Journal of Zhengzhou University (Science Edition), 2019, 51(2):93-96
[4] Zhang YJ, Bai L. A fast symbolic data clustering algorithm based on symbolic relationship graph. Computer Science, 2021, 048(004):111-116.
[5] Shao CL, Sun TF, Ding SF. Clustering integration algorithm based on information entropy weighting. Journal of Nanjing University (Natural Science Edition). 2021, 57(2):189–196.
[6] Zhang YJ. Symbolic relation graph construction and cluster analysis for symbolic data. Shanxi University, 2021. DOI: 10.27284/d.cnki.gsxiu.2021.001815.
[7] El-Sonbaty Y, Ismail MA. Fuzzy clustering for symbolic data. IEEE Transactions on Fuzzy Systems, 1998, 6(2):195-204.
[8] Fishe R, Douglas H. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 1987,2(2):139-172.
[9] Yao LY, Qian XZ, Fan L. Incremental clustering algorithm based on cluster features. Sensors and Microsystems, 2019, 38 (1): 3. DOI: CNKI: SUN: CGQJ.0.2019-01-043.
[10] Ralambondrainy H. A conceptual version of the K-means algorithm. Pattern Recognition Letters, 1995, 16(11):1147-1157.
[11] Barbará D, Li Y, Couto J. COOLCAT: an entropy-based algorithm for categorical clustering. International Conference on Information and Knowledge Management. 2002: 582-589.
[12] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304.
[13] Dinesh MS, Gowda KC, Nagabhushan P. Unsuper-vised classification for remotely sensed data using fuzzy set theory. Geoscience and Remote Sensing (IGARSS'97). IEEE Press,1997.
[14] Sudipto G, Rajeev R, Kyuseok S. Rock: A robust clustering algorithm for categorical attributes. Information Systems, 2005(5): 345-366.
[15] Sharma S, Singh M. Generalized similarity measure for categorical data clustering. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, Press, 2016:765-769.
[16] Nguyen THT, Huynh VN. A k-means-like algorithm for clustering categorical data using an information theoretic-based dissimilarity measure. International Symposium on Foundations of Information & Knowledge Systems. Springer-Verlag New York, 2016.
[17] Ding X, Tan J, Wang M. A categorical data clustering algorithm and its efficient parallel implementation// 2016 5th International Conference on Computer Science and Network Technology (ICCSNT). IEEE Press,2017:224-228.
[18] Jia B, Liang Y, Su H. An improved K-Modes clustering algorithm. Software Guide, 2019,18(6):60-64.
[19] Wang ZH, Liu ST, Luo Q. KNN Classification Algorithm based on improved K-modes clustering Computer Engineering and Design,2019(8):2228-2234.
[20] Mahamadou AJD, Antoine V, Christie G J, et al. Evidential clustering for categorical data// 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE Press,2019:1-6.
[21] Huo GQ, Lu JB, Luo SX. Research on image mosaic based on CLAHE and improved ZnCC. Progress in Laser and Optoelectronics, 2022, 59(12):9.
[22] Niu YK. Research on JPEG image forensics technology. Beijing: Beijing Jiaotong University, 2021.
[23] Wei JX. Research on interdisciplinary knowledge discovery and visualization. Nanjing University, Nanjing, 2010.
[24] McDaid AF, Greene D, Hurley N. Normalized Mutual information to evaluate overlapping community finding algorithms[J]. arXiv:1110.2515.
[25] Warrens MJ. On the equivalence of Cohen's kappa and the hubertarabie adjusted rand index. Journal of Classification, 2008, 25(2): 177-183.
[26] Yang RL. Research on Web document clustering method based on phrase features. Chongqing University, 2010.