NOVEL EVALUATION INDEX OF CROSS-SCALE DISCRETIZATION UNCERTAINTY BASED ON LOCAL STANDARD SCORE

Jiangping Chen [1, *], Wanshu Feng [1], Yan Huang [1]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei Province, China - chen_jp@whu.edu.cn; 1253787248@qq.com; 2017282130202@whu.edu.cn

Commission IV, WG IV/3

KEY WORDS: Discretization; Uncertainty; Local Standard Score; Individual Assessment

ABSTRACT:

Optimal discretization of continuously valued attributes is an uncertainty problem. The uncertainty of discretization is propagated and accumulated in the process of data mining, which has a direct influence on the usability and operation of the output results for mining. To address the limitations of existing discretization evaluation indices in describing accuracy and operation efficiency, this work suggests a discretization uncertainty index based on individuals. This method takes the local standard score as the general similarity measure in and between the intervals and evaluates discretization reliability according to the relative position of individuals in each interval. The experiment shows the new evaluation index is consistent with commonly used metrics. Under the premise of guaranteeing the validity of discrete evaluation, the proposed method has greater description accuracy and operation efficiency than extant approaches; it also has more advantages for massive data processing and special distribution detection.

1. INTRODUCTION

The discretization of continuously valued attributes is an essential and important step during data mining. Many existing data mining algorithms only target discrete attributes and when these algorithms are applied, the continuous attributes are first discretized. There is a plethora research on discretization effect evaluation, and many classical evaluation coefficients have also been widely used. According to different evaluation objects, the existing evaluation of discretization effects can be divided into two levels: global evaluation and individual-based evaluation. The global evaluation method evaluates the overall rationality of the discretization interval. The commonly used coefficients, such as the Dunn validity coefficient proposed by Dunn et al., calculate the ratio of the maximum distance in the interval to the minimum distance between the intervals. Superior discretization should have a large distance between the intervals, and the cohesion in the interval is strong. Chou et al. proposed the CH coefficient, and evaluated the discretization effect by calculating the sum of the squared distances of the data distances in the interval and the square of the distance between the center points of the entire data set. Similarly, the I coefficient proposed by Davies and Bouuldin also evaluates discretization by calculating the distance within the interval and between the interval. However, the global coefficient quantitatively evaluates only one coefficient value for a column of data, and it is impossible to describe and analyze the discretized structural details. Therefore, individual-based assessment can effectively compensate for this deficiency. For example, Rousseeuw raised the Silhouette Index to calculate the coefficient values for each object in the data set, and evaluate the discretization superiority of each object. However, facing the large data sets, the method of Silhouette Index requires more time to perform multiple calculate on an object-by-object basis. With the rapid increase in data size and the increasing complexity of data forms, there is an urgent need for new discretization assessment methods that adapt to massive data sets and complex data patterns.

Therefore, a new discretization uncertainty coefficient evaluation method based on local standard score construction was proposed in this paper. This method uses the standard score to construct the overall similarity measure of the data within the interval and between the intervals, and evaluates the discrete reliability by considering the relative position distribution of the individual in each interval.

1.1 The source of continuous attribute discretization uncertainty

In the process of discretization of continuous attributes, the distribution characteristics of the original data set will be changed due to the concept level of the attribute. Mapping continuous data to several discrete intervals loses the continuous change details of complex data, and the amount of information contained in the data is also reduced, resulting in uncertainty in discretization results and application analysis. The uncertainty description in the discretization result consists of two parts: the discretization term set X and the probability distribution $U(X) \in [0, 1]$ of this set. For a data set containing m consecutive attributes, the number of data entries in the attribute value range is recorded as n. The uncertainty of the discretization interval corresponding to any record X under each successive attribute is represented by a discrete interval value representing the attribute i in the record j, U_{ij} indicates its corresponding uncertainty.

Therefore, the uncertainty in discretization is a kind of uncertainty derived from the data itself and the concept. Each record in the data set (each attribute value field) can be represented by the corresponding uncertainty probability. The uncertainty can be divided into two aspects: the degree of cohesion within the interval (reflecting how closely the objects in the interval are closely related) and the degree of separation

* Corresponding author

between the segments (reflecting where a certain interval is different from other intervals).

Discretized cohesion and resolution are manifested in this relationship. The effectiveness evaluation of the results of discretization alone with cohesion or resolution is unreliable. The more division intervals there are, the greater the degree of separation between the intervals, and the smaller the degree of cohesion within the interval. Under extreme conditions, each object corresponds to a subinterval, with itself as the center of the interval. At this time, the error in the interval is 0, and there is no uncertainty, but the ideal discretization is actually not discretized, which is not conducive to subsequent calculation and analysis. Therefore, the discretization uncertainty assessment needs to be based on a combination of cohesion and resolution, and finding a balance between the cohesion within the interval and the resolution between the intervals, so as to evaluate the discretization results.

1.2 Discretization uncertainty coefficient

During an individual-based uncertainty assessment, each object in the discretization result is calculated for its uncertainty in the discrete interval to which it belongs. The concept of standard scores is introduced in this study to evaluate the degree of cohesion in the interval and the degree of separation between intervals.

The standard score is the measure of the discrete distribution between the object and the mean in units of standard deviation. Firstly, compare the original value of an object in the discrete interval with the average level of the interval, and then judge the continuous distribution level of the object from the whole by the standard deviation. It can be seen that the standard score is a quantification of the distribution level of an object in this discrete interval. Compared with the average distance, the calculation of the standard score can reflect the relative standard distance of the object distance interval or the neighborhood interval, so that the deviation distribution level of the individual object in the discretization can be better evaluated. The individualized degree of cohesion is measured by calculating the standard score of each object's distance from the center of the interval in the interval based on the individual's uncertainty coefficient, and the individual score is measured using the standard score of each object from the center of the nearest neighbor. The value of the coefficient is obtained by the ratio of the degree of aggregation to the degree of separation.

For a given continuous attribute $X=\{x_1,x_2,...,x_n\}$, n is the number of data objects in X. Using some discretization algorithm to divide X into k intervals, the discretization result is recorded as $I_1,I_2,...,I_k$. For the i[th] object $x_i(x_i \in I_j, i \in [1,n], j \in [1,k])$, the uncertainty coefficient is:

$$ui_i = \frac{a_i}{Max(a_i,b_i)} \tag{1}$$

$$a_i = \frac{d(x_i,m_j)}{std(I_j)} \tag{2}$$

$$b_i = \begin{cases} \dfrac{d(x_i,m_{j-1})}{std(I_{j-1})} & x_i \le m_j \cup j=k \\ \dfrac{d(x_i,m_{j+1})}{std(I_{j+1})} & x_i > m_j \cup j=1 \end{cases} \tag{3}$$

Where a_i is the standard score of x_i relative to its associated interval I_j, and b_i is the standard score of x_i relative to its neighborhood interval I_{j-1}, I_{j+1}. If I_j is the boundary interval, the single adjacent interval is only taken. The uncertainty coefficients UI_k for discrete intervals are defined as follows:

$$UI_k = \frac{1}{n}\sum_{i=1}^{n} u_i \tag{4}$$

Where n is the number of objects in the data set, and k (the average uncertainty coefficient) is the number of discrete intervals, which is a measure of the overall uncertainty of the partitioning of the entire data set. The value of the uncertainty coefficient varies from 0 to 1. A value of 1 indicates that the discrete distribution of the object relative to its interval is equal to or greater than the degree of dispersion of the relative neighborhood interval, and the uncertainty reaches the highest. The closer the value is to 0, the smaller the uncertainty of the object.

1.3 Comparison of discretization evaluation coefficients and characteristics of uncertain coefficients

There have been many results on the evaluation of discretization effects as well as on classical evaluation coefficients. Although the starting points of these two methods are different, they are both based on the overall similarity of the data between the sub-intervals and sub-intervals in the discretization results. The distance of the data represents the degree of difference between the data. According to the discretization requirement, the result of the discretization of the continuous attribute should make the distance of the data between the intervals as large as possible, and the distance of the data in the interval smaller. According to different evaluation objects, the existing discriminant effect evaluation can be divided into two levels: global evaluation and individual-based evaluation. The existing mature evaluation coefficient is compared with the construction principle of the uncertainty coefficient proposed in this study. The results are shown in Table 1 below.

Compared with the existing discretization evaluation coefficient, the uncertainty coefficient proposed in this paper has the following characteristics.

(1) Consider the contribution of each individual's discretization to uncertainty.

The uncertainty coefficient is calculated in units of individuals, which fully reflects the heterogeneity of the discrete individuals. The average uncertainty coefficient is obtained from the discretization uncertainty of each individual. Compared with the traditional global evaluation coefficient, it has a significant improvement in the granularity and flexibility reflecting the

	Formula	Cohesion	Resolution	Evaluation Unit
Uncertainty coefficient	$u_i = \frac{a_i}{Max(a_i, b_i)} \ / \ U_k = \frac{1}{n}\sum_{i=1}^{n} u_i$	The standard score of the individual from the center of the interval	Standard score of the individual from the center of the nearest neighbor	Individual/global
Contour coefficient	$s_i = \frac{n_i - m_i}{Max(m_i, n_i)} \ / \ S_k = \frac{1}{n}\sum_{i=1}^{n} s_i$	Average distance of individuals to other samples in their interval	Average distance of all objects from the individual to the nearest neighbor	Individual/global
Dunn coefficient	$D(k) = \min_{1 \le i \le k}\{\min_{i \ne j} \frac{\min_{x \in I_i, y \in I_j} d(x,y)}{\max_{x,y \in I_k} d(x,y)}\}$	The range of the largest interval in all intervals	The smallest two-point distance between the interval and the interval	Global
Calinski-Harabasz (CH) coefficient	$CH(k) = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i d^2(m_i, m)}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{x \in I_i} d^2(x, m_i)}$	The sum of the squares of the distances between the points in the interval and the center of the interval	The sum of the squares of the distance between the center of each interval and the center of the data set	Global
I coefficient	$I(k) = \left[\frac{1}{k}\frac{\sum_{x \in C} d(x,m)}{\sum_{i=1}^{k}\sum_{x \in I_i} d(x, m_i)} \max d(m_i, m_j)\right]^p$	The sum of the distances from the points in the interval to the center of the interval	The maximum distance between the interval and the center distance of the interval	Global

Table 1. Comparison of discretization evaluation coefficient

degree of discretization reliability. Therefore, extensive analysis of the refined uncertainty distribution can be explored.

(2) Efficiency improvement

Compared to the current individual-based contour coefficient, the uncertainty coefficient significantly improves operational efficiency. The contour coefficient needs to scan the data repeatedly during the calculation. However, the calculation of the uncertainty coefficient is based on the standard score as the dimension of the degree of cohesion and resolution, which greatly simplifies the time complexity of the coefficient calculation, thereby improving the evaluation efficiency and better service to the analysis and application of massive data.

2. METHODOLOGY

2.1 Discretization uncertainty coefficient verification

2.1.1 Effectiveness verification

To verify whether the uncertainty coefficient proposed in this study can effectively evaluate the reliability of the discretization results, this paper compares the uncertainty coefficient with the existing evaluation coefficient in the global evaluation and individual evaluation. The experimental data includes one set of simulation data and one set of actual data. The experimental hardware comprises an Intel Core i7 with 3.60GHz CPU and 8GB memory. The operating system platform is Microsoft Windows 7 Ultimate, and the software programs are Microsoft Visual C++ 6.0 compiler and Matlab R2014a.

The experimental data verified by the global evaluation is simulated data and contains 500 samples. Four range partitions were added to the continuously evenly distributed data to form five separate intervals, as shown in Figure 1. The experimental data known for discrete distributions were discretized into 2-10 classes using EW, K-means and FCM algorithms. In addition to the uncertainty coefficient U, the global evaluation coefficient (Dunn coefficient, CH coefficient, I coefficient) is used to evaluate the discretization effect, so as to verify the validity of the uncertainty coefficient applied to the discretization evaluation. The relevant data is shown in Table 2.

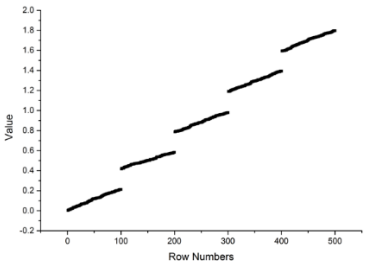


Figure 1. Distribution of simulated data

	EW				K-means				FCM			
	Dunn	CH	I	U	Dunn	CH	I	U	Dunn	CH	I	U
2	0.01	1443.69	0.92	0.30	0.02	1488.64	0.98	0.258	0.01	1450.02	0.92	0.30
3	0.01	1991.61	1.29	0.33	0.01	2006.55	1.51	0.330	0.00	1916.32	1.38	0.32
4	0.01	2250.22	2.04	0.36	*0.02*	2687.70	2.94	0.301	0.01	2702.74	3.00	0.29
5	0.95	11486.75	9.40	0.14	0.01	2081.66	2.24	0.297	0.95	11486.75	9.40	0.14
6	0.02	9627.19	7.60	0.18	0.03	11102.38	8.87	0.160	0.03	11041.21	8.73	0.16
7	0.03	10282.79	7.24	0.15	0.03	9600.77	7.21	0.164	0.01	11209.40	7.59	0.17
8	0.02	9599.05	5.84	0.14	0.00	7361.49	5.62	0.169	0.01	10191.86	7.00	0.17
9	0.00	6739.45	3.64	0.16	0.03	7289.95	4.77	0.166	0.02	10575.11	7.03	0.20
10	0.01	9606.84	6.45	0.17	0.02	6470.23	3.91	0.167	0.01	10314.64	7.96	0.21

Table 2. Effectiveness evaluation of simulated data under three discretization algorithms

Considering the dimensional difference between the coefficients, the four series of values obtained by the four coefficients are normalized, so that the evaluation coefficients are applied to the simulation data and the effects of discretization into 2-10 intervals under different discretization methods are evaluated and compared. Specifically, the method is to subtract the minimum value in the column from the calculated value of each coefficient evaluation, and then divide by the difference between the maximum value and the minimum value of the data in the series. For the value trend of the uncertainty coefficient U and the existing coefficient (U=1-U), the larger the value, the more reasonable the discretization. Figure 2 shows the comparison between the results of the four coefficient evaluations after standardization.

Since the metrics for cohesion and resolution vary, the calculation results of the existing evaluation coefficients are different. The evaluation of the discretization results can be divided into two aspects: the detection of the optimal number of discrete intervals and the comparison of the discretization methods. Take Figure 2 (a) as an example, the EW method is used to discretize the simulation data into two-ten intervals. The four evaluation coefficients show that the discretization is most reasonable when the simulation data is divided into five discrete intervals of equal width. This is consistent with known discrete distribution characteristics of data. The uncertainty coefficient U proposed in this study is consistent with the calculation results of the current classical coefficients in the detection of the optimal number of discrete intervals.

By comparing the optimal number of simulation data under the three discretization methods of EW, K-means, and FCM and the values of the four evaluation coefficients, the following can be known.

（1）For the EW method and the FCM method, the four coefficients show that the best discretization effect is achieved when divided into five discrete intervals, and the interval range is completely the same; that is, the simulated data is divided into five equal parts with the same volume. K-means is best when segmenting the simulated data into six intervals.

（2）Comparing the coefficient values of the three discretization methods under the optimal interval number, it can be seen that the values of the Dunn coefficient, the CH coefficient and the I coefficient of the K-means method are lower than the EW method and the FCM method, but the coefficient U is not determined. Figure 2 (b) shows that the coefficients in the corresponding graphs of the EW method and the FCM method are the normalized maximum value 1 at the 5th interval. However, the K-means method reaches the series maximum at the 6th interval, but both are less than 1, and the Dunn coefficient is even 0.03,

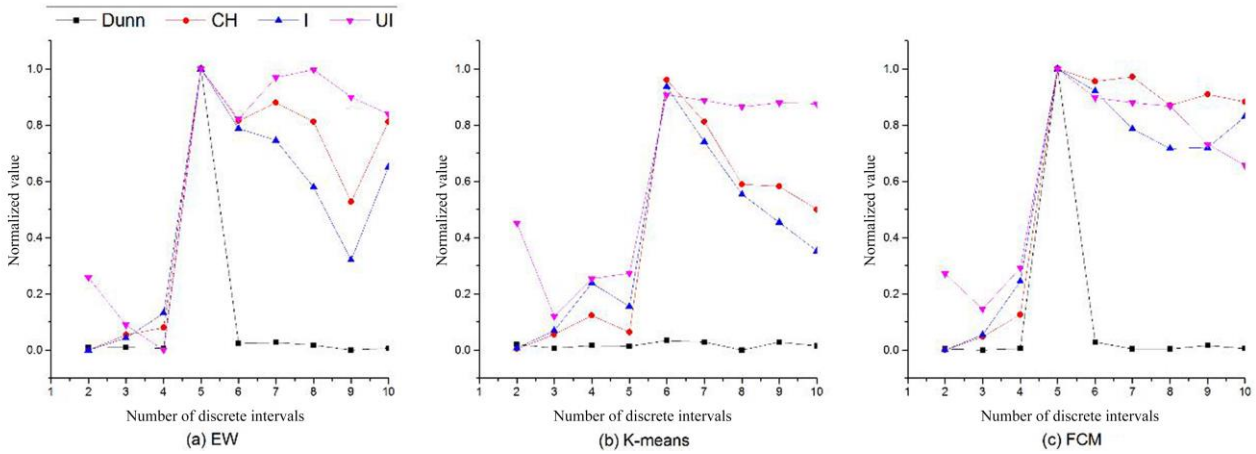


Figure 2. Evaluation of discrete effect of simulated data

whose discretization effect is worse than the other two methods. Therefore, the discretization of the simulated data should be divided into five discrete intervals using the EW or the FCM method, which is consistent with the known data distribution characteristics.

Through the experimental evaluation and comparison of the known discrete distribution simulation data, it can be seen that in the comparative evaluation of the detection and discretization methods of the optimal discrete interval number, the uncertainty coefficient U proposed in this study is compared with the evaluation result of the current coefficient. Consistency with the actual distribution characteristics of the data is an effective way to evaluate the results of discretization.

2.1.2 Individual evaluation

The individual evaluation verification uses the IRIS data set to compare the existing discretization evaluation methods and the evaluation results of the uncertainty coefficients. IRIS data, also known as the iris flower dataset, is recognized as the most famous dataset for data mining. The data set consists of 150 data consisting of four consecutive attributes, which are the length of the flowerbed, the width of the flower, the length of the petals, and the width of the petals.

In the experiment, the K-means algorithm is applied to discretize the four consecutive attributes in the IRIS data set, which are divided into three categories for comparison. A comparison of discretization evaluation coefficients (see section 2.3) reveals that only the contour coefficients and the uncertainty coefficients proposed in this study can evaluate the discretization uncertainty of each data, and verify the validity of the proposed uncertainty coefficient applied to discretization evaluation. The contour coefficient calculates the average distance to other individuals in the dataset and to all individuals in the most adjacent interval for each individual in the data set, thereby jointly evaluating the discretization superiority of each individual.

The K-means algorithm is used to discretize the four attributes respectively to obtain the S and U coefficients of each individual, then the individual uncertainty evaluation is conducted. Since the contour coefficient S and the uncertainty coefficient U are opposite in the evaluation reliability, the difference between the effective analogy uncertainty coefficient U and the current evaluation method is that U is taken as 1-U; if the value is larger the discretization is more reasonable. The results of comparison of the similarity between S and U during the individual evaluation are shown in Figure 3. The abscissa indicates the number of data records, and the ordinate indicates the calculated contour coefficient S and the uncertainty coefficient U proposed in this study.

The four consecutive attributes in the IRIS data set, the S and U coefficients can effectively reflect the uncertainty of each of the data in the discretization. For each discrete interval, the individual discretization reliability in the middle of the interval is higher, and the individual's uncertainty increases when it is closer to the segmentation point. Furthermore, it is understood that the contour coefficient S and the uncertainty coefficient U each have a strong similarity of 0.8 or more in each experimental data when the Pearson correlation of the S and U coefficient sequences are calculated. During the individual-based discretization uncertainty assessment, the uncertainty coefficient U based on the local annotation score and the existing contour coefficient S have an extremely high distribution similarity, and the discriminant uncertainty evaluation effect for each individual is consistent.

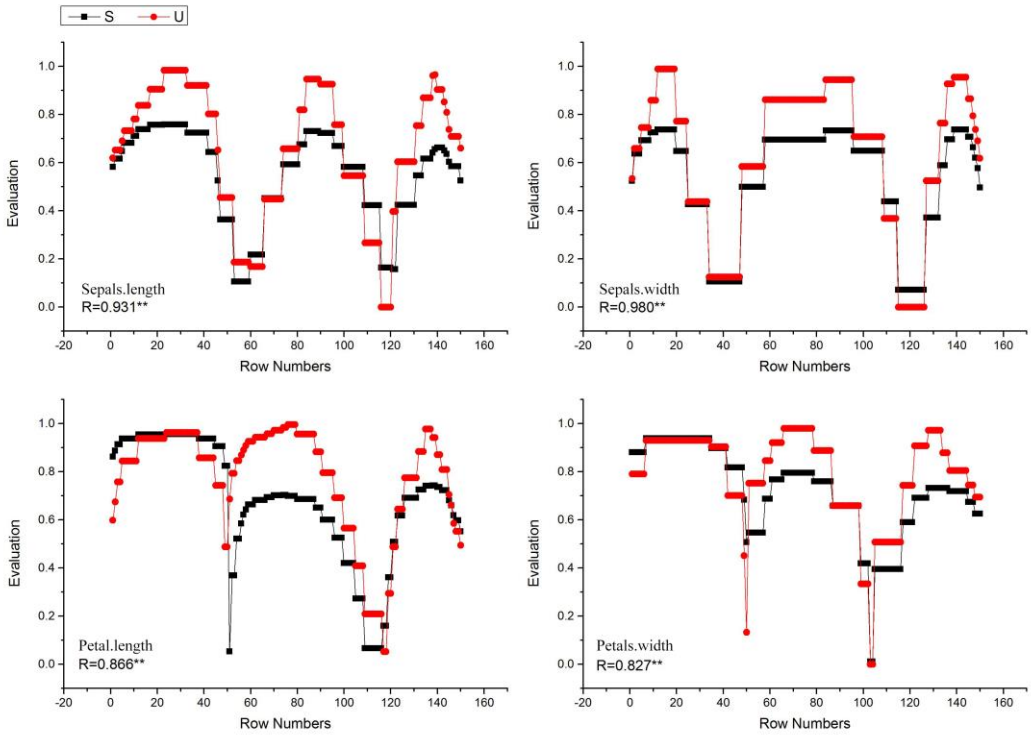


Figure 3. Comparison of individual evaluation

2.2 Superiority verification

2.2.1 Calculation efficiency

It can be seen from the experiment in section 3.1 that the proposed uncertainty coefficient U is basically consistent with the rule mining result obtained by evaluating the discretization reliability of the existing contour coefficient S. For the uncertainty coefficient method, since it simplifies the time

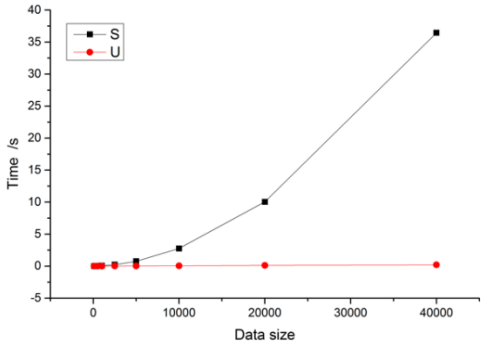


Figure 4. Running-time experiment of Silhouette coefficient and uncertainty coefficient

complexity of the algorithm, it will achieve higher processing efficiency in the face of massive data. Take analog data as an example, a random number of normal distributions are generated in Matlab, and the amount of data is gradually increased from 100 to 40,000. The data set is discretized into three intervals by the K-means algorithm. In Matlab, the validity of each simulation data set is verified by the contour coefficient S and the uncertainty coefficient U. As the amount of data increases, the calculation time is as shown in Table 3. It can be seen that with the increase of data volume, the uncertainty coefficient U proposed in this study has significant time superiority as shown in Figure 4, which greatly reduces the time consumption of individual evaluation and improves the efficiency of data processing.

Data size	100	1000	5000	10000	20000	40000
S /s	0.016	0.08	0.769	2.76	10.04	36.45
U /s	0.012	0.016	0.04	0.057	0.131	0.205

Table 3. Comparison of Silhouette coefficient and uncertainty coefficient

2.2.2 Breakpoint recognition

Firstly, a random data set with 1000 entries is generated by Matlab2016. Then two partition points are set to the value domain partition to form three separate intervals, and the simulated dataset of the hierarchical sequence distribution is obtained. The K-means method is used to discretize this dataset into low, medium and high, which are labeled 1 to 3, and their distribution is shown in Figure 5 (a). It can be seen that the 1st interval includes a portion with a range of 0-0.1 and 0.4-0.64, which is partitioned across a range. The 2nd interval range is 0.64-1.15, which is evenly distributed within the interval. The 3rd interval range is 1.55-1.7, which forms a value domain fault with the 2nd interval.

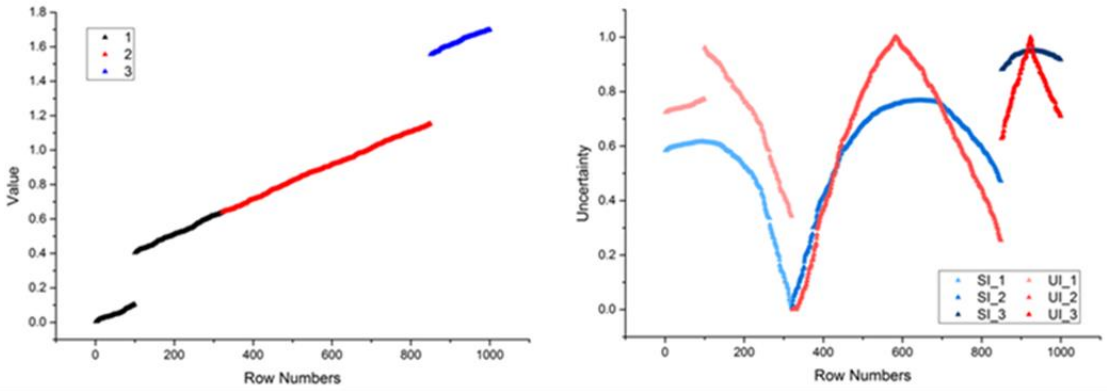
The uncertainty of the simulated data is evaluated by the contour coefficient S and the uncertainty coefficient U, and the results are shown in Figure 5 (b) (U takes 1-U). It can be seen that the overall distribution of S and U is relatively consistent, the reliability is higher in the middle of each interval, and that the uncertainty is closer to the boundary area.

Since the distribution of the 1st interval spans a range of faults, the fault of the uncertainty coefficient U also appears as a reliability fault, and the data of the value range of 0-0.1 is higher than the data of the other side of the fault in terms of interval separation; the uncertainty is therefore stronger. However, the data on the right side of the partition is better than the majority of the data in the interval, which shows high reliability. In the evaluation of U, the standard score is selected as the unit of distance metric, and its value reflects the relative position of the individual in the interval, so it can better reflect the uncertain distribution of data faults in this interval. The S-evaluation uses the average distance as the metric, whose uncertainty evaluation is a smooth result, and the intra-segment differentiation cannot be detected. Therefore, in the special distribution with data faults, the uncertainty coefficient can better evaluate the deviation distribution level of individual objects in discretization.

3. CONCLUSIONS

In this study, a new cross-scale discretization uncertainty measure coefficient based on local standard score construction is proposed, which realizes the controllable experimental analysis of the comprehensive performance of measurement quality and computational efficiency of discretization uncertainty. Considering the individual's comprehensive contribution rate within and between discrete intervals, the distribution of discretized uncertainty is detected and evaluated. The experimental results show that the evaluation effect on the discretization reliability is consistent with the existing commonly used evaluation coefficients. Comparing the uncertainty coefficient and the existing evaluation coefficient to evaluate the operation efficiency of the discretization effect of large-volume data, we can see that the uncertainty coefficient proposed in this study significantly shortens the calculation time, and the calculation efficiency of massive data is better than the existing evaluation coefficient. In addition, the uncertainty coefficient helps to identify breakpoints and abrupt points in the dataset, which is more suitable for discretization evaluation of special distributions. Furthermore, the value normalization of the uncertainty coefficients constructed using the local standard scores varies from 0 to 1, and such evaluation results directly support the unified comparative analysis of the discrete degrees of uncertainty of different types of attributes. Potentially, the statistic of the standard fractional distribution of the standard normal distribution N (0,1) (discrete uncertainty coefficient) can develop a probabilistic theoretical analysis of the nature of the statistical (estimated) amount of data.

Since discretization results are unlikely to form a one-to-one ideal mapping for the complex real world, all types of discretization bring some uncertainty. Furthermore, when the discretization results are applied to data mining, and the



(a) Discretization of experimental data (b) Comparison of evaluation results

Figure 5. Verification of superiority in description of uncertainty coefficient

uncertainty of the previous stage is propagated to the latter stage, the result is the accumulation and propagation of uncertainty. Therefore, it is of great value to evaluate the individual uncertainty in discretization. Although some preliminary explorations on evaluating the individual uncertainty in discretization has been carried out in this study, there are still many areas for improvement in follow-up research.

Discretization uncertainty assessment can be effectively applied to discretization algorithms and interval number selection of actual data.

ACKNOWLEDGMENTS

This study was supported by a grant from the National Natural Science Foundation of China (project No: 41331175).

REFERENCE

Alataş B, Akin E. An Efficient Genetic Algorithm for Automated Mining of Both Positive and Negative Quantitative Association Rules[J]. Soft Computing, 2006, 10（3）: 230-237.

Augasta M G, Kathirvalavakumar T. A New Discretization Algorithm Based on Range Coefficient of Dispersion and Skewness for Neural Networks Classifier[J]. Applied Soft Computing, 2012, 12（2）: 619-625.

Aumann Y, Lindell Y. A Statistical Theory for Quantitative Association Rules[J]. Journal of Intelligent Information Systems, 2003, 20（3）: 255-283.

Berry M J, Linoff G. Data Mining Techniques: For Marketing, Sales, and Customer Support[J]. John Wiley & Sons,Inc, 1997, 17（1）: 45-53.

Bezdek, James C. Pattern Recognition with Fuzzy Objective Function Algorithms[J]. Advanced Applications in Pattern Recognition, 1999, 22（1171）: 203-239.

C.Bezdekâ€ J. Cluster Validity with Fuzzy Sets[J]. Journal of Cybernetics, 1974, 3（3）: 58-73.

Chaves R, Ramírez J, Górriz J M. Integrating Discretization and Association Rule-Based Classification for Alzheimer's Disease Diagnosis[J]. Expert Systems With Applications, 2013, 40（5）: 1571-1578.

Chou C H, Su M C, Lai E. A New Cluster Validity Measure and Its Application to Image Compression[J]. Pattern Analysis & Applications, 2004, 7(2): 205-220.

Davies D L, Bouldin D W. A Cluster Separation Measure[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1979, 1(2): 224-227.

Dunn J C. Well-Separated Clusters and Optimal Fuzzy Partitions[J]. Journal of Cybernetics, 1974, 4(1): 95-104.

Elkano M, Galar M, Sanz J A, et al. Consensus Via Penalty Functions for Decision Making in Ensembles in Fuzzy Rule-Based Classification Systems[J]. Applied Soft Computing, 2017:

Hauke J, Kossowski T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data[J]. Quaestiones Geographicae, 2011, 30（2）: 87-93.

Hong J, Qin M. Multisymplecticity of the Centred Box Discretization for Hamiltonian Pdes with M ≥ 2 Space Dimensions[J]. Applied Mathematics Letters, 2002, 15（8）: 1005-1011.

Kaufman L, Rousseeuw P J. Finding Groups in Data. An Introduction to Cluster Analysis[M]//DBLP. 2005.

Krishnamoorthy S, Sadasivam G S, Rajalakshmi M, et al. Privacy Preserving Fuzzy Association Rule Mining in Data Clusters Using Particle Swarm Optimization[J]. International Journal of Intelligent Information Technologies (IJIIT), 2017, 13(2): 1-20.

Kumar V, Chhabra J K, Kumar D. Performance Evaluation of Line Symmetry-Based Validity Indices on Clustering Algorithms[J]. Journal of Intelligent Systems, 2016, 26(3): 254-267.

Mcfeeters S K. The Use of the Normalized Difference Water Index （Ndwi） in the Delineation of Open Water Features[J]. International Journal of Remote Sensing, 1996, 17（7）: 1425-1432.

Mennis J, Guo D. Spatial Data Mining and Geographic Knowledge Discovery—an Introduction[J]. Computers Environment & Urban Systems, 2009, 33(6): 403-408.

M M-B, F M-, A T, et al. An Evolutionary Algorithm to Discover Quantitative Association Rules in Multidimensional Time Series[J]. Soft Computing, 2011, 15（10）: 2065-2084.

Nishida T Y. An Approximate Algorithm for Np-Complete Optimization Problems Exploiting P Systems[C]//proceedings of the Proceedings of Brainstorming Workshop on Uncertainty in Membrane Computing, Palma de Mallorca,Espaa, F, 2004: 185-192.

Rousseeuw P. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20(20): 53-65.

Steinbach, Michael. Introduction to Data Mining[M]. China Machine Press, 2010.

Tao Gang, Yan Yonggang, Liu Jun, et al. Discrimination of continuous attributes based on improved SOM clustering[J]. Journal of Computer Applications, 2015, 35(S1): 89-92.

Wang L. Discretization Algorithm of Rough Set Continuous Attributes Based on Improved Particle Swarm Optimization[J]. Computer Engineering & Applications, 2010, 46（15）: 115－118.

Wang Luxin. Discretization MethodS of Attributes of Power Big Data based on Cloud Computing Technology[J]. Digital Technology and Application, 2015, (1): 56-58.

Xie Juanying, Zhou Ying, Wang Mingzhao, et al. The new criteria for evaluating clusterings of clustering algorithms[J]. Transactions on Intelligent Systems, 2017, 14(6): 1-9.

YangYan, Jin Fan, Mohamed K. Survey of Clustering Validity Evaluation[J]. Application Research of Computers, 2008, 25(6): 1630-1633.

Zhang Yusha, Jiang Shengyi. Suevey on Continuous Features Discretisation Algorithm[J].Computer Application and Software, 2014, 31(8): 6-8.

Zhou Kaile, Yang Shanlin, Ding Shuai. A Review of Cluster Validation[J]. Systems Engineering-Theory and Practice, 2014, 34(9): 2417-2431.

Zhu Lianjiang, Ma Bingxian, Zhao Xuequan. Clustering Validity Analysis based on Silhouette Coefficient[J]. Journal of Computer Applications, 2010, 30(S2): 139-142.