ICIC Express Letters
Part B: Applications
Volume **8**, Number **10**, October **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 1355–1362

DISCRETIZATION ALGORITHM BASED ON THE MINIMUM GINI INDEX AND OPTIMIZATION ON FORMAL CONTEXT

Jialin Song, Huaixin Liang and Wenxue Hong*

Institute of Electrical Engineering
Yanshan University
No. 438, Hebei Avenue, Qinhuangdao 066004, P. R. China
*Corresponding author: hongwx@ysu.edu.cn

Received April 2017; accepted June 2017

Abstract. *The Formal Concept Analysis (FCA) is one of the essential tools to data mining and knowledge discovery; however, the formal context needed during the data processing is binary. Therefore, finding a proper and effective method to scatter the data is playing an essential role to the precision of pattern classification and the generation of formal context. In this paper, we proposed a discretization algorithm based on the minimum Gini index and the optimization on formal context, which realized the function that deals with continuous data in information system and extracts the formal context that can be used in FCA. Firstly, the Gini index of each potential split point is defined, on which the rule of importance to the split points is based. Then the formal contexts are sent to the random forest classifier to calculate the mean error rate. Meanwhile, we compared the method we proposed with other two methods by using the standard data from UCI database. The result shows that the precision of the method we proposed is better than that of the other two methods generally. It is proven that the algorithm is effective.*

Keywords: FCA, Gini index, Optimization algorithm, Discretization

1. **Introduction.** German professor Wille proposed Formal Concept Analysis (FCA) firstly in 1982, which is applied widely on mining on concepts, knowledge discovery [2,3], machine learning [4], software engineering [5], information retrieval [6], visualization [7] and other essential fields. The FCA analyzed data by extracting the useful rules from complex formal context and by setting up the hierarchy between objects and attributes. Basically, the formal context is binary which means finding an effective granulation algorithm to deal with a number of continuous data is of great essence.

The classic rough set theory, a kind of mathematical tool for the uncertain blurry knowledge, was proposed by Poland professor Pawlak, which is widely used in many areas like pattern recognition, knowledge discovery, and fault detecting. During the process of data analysis of rough set, the discretization of data is supposed to be done if the type of condition attribute or the decision attributes are continuous. According to the researches done by many scholars and to whether or not the class information is used, the discretization can be defined as supervised discretization and unsupervised discretization. The equal frequency method and equalization method, known as easily operated and understood, are the typical unsupervised discretization. However, the precision can be affected because of the ignorance to the essential class information. Besides, Usama and Irani [8] proposed a discretization method based on information entropy. [9,10] proposed some optimized greedy algorithms based on genetic algorithm. [11] discussed the discretization method based on the set of optimal split points on GA considering the fact that the split points set can largely affect the process precision. [12] completed the discretization by introducing a cloud model into the data processing. [13] also increased the precision by using the optimized method based on the information entropy. Besides, [14] proposed a new method based on the visualization method, presenting the data distribution visualization, converted the data distribution to figure distribution and fuzzy analysis to form the formal context.

In this paper, we proposed a new discretization method by combining the minimum Gini index and optimization to the formal context. Firstly, calculation on the Gini index of each potential split point of continuous data is made. Then, the data of minimum Gini index is chosen to be the potential split point. The process is finished by dividing the whole object domain, which means the split points chosen can cover all the objects. The optimization of the formal context based on the clustering theory and the new method on exchanging the rows and columns is proposed to get the purer granules. The method can be used to extract the formal context that can be used in formal concept analysis from the information system. Finally, a pattern recognition system is designed to testify the effect of combination result of our method and other two methods based on UCI database. The result shows that the precision of the method we proposed is all higher than that of the other two methods, especially, the results on database of Sonar, Glass and Wine are better eminently. Best results of the mean error rate are capable of reaching 0%.

2. **Basic Theory.**

Definition 2.1. *The information sheet $K = \{P, M, G\}$, a special representation of the information system, is called in the Formal Concept Analysis a formal context, which concludes the objects set P, attributes set M and the binary relationship G, where $G \subseteq P \times M$. The rows of formal context point to the objects set, and the columns point to attributes. Supposed that one object had one certain attribute, the sign "1" should be put at the intersection of them; otherwise, set "0" instead. The formal context based on theory above is shown in Table 1. In the formal context, the number of one row represents an object and the letter of one column means one attribute.*

The granule is the basic unit, on which the granular computing is based. Granules in different layers contain relations of many kinds, in which the granule is the prior objects of the layers concerned. The essence of the discretization of continuous data is actually a process of granulation, which is also the basic part of granular computing, and is a process of generating granules. When the principle of granulation is given, the granule layers concerned that are the ingredients of the whole granule structure are supposed to be gotten.

Table 1. Formal context

	a	b	c	d
1	1	1	0	0
2	1	1	0	1
3	1	0	1	0

Definition 2.2. *An information system can be defined as $S =< U, R, V, f >$, where $U = \{x_1, x_2, \dots, x_n\}$ represents the domain of the non-empty limited set that is composed of many objects. $R = C \cup \{d\}$ represents the attribute set, where C is the conditional attribute set and $\{d\}$ is the decision set. $V = \cup\{V_r|r \in R\}$, where V_r is the range of attribute r, $f : U \times R \to V$ is the information function and $\forall r \in R$, $x \in U$, $f(x, r) \in V_r$. Provided that the number of the classes is $r(d)$, the attribute set (a, c) will be a split interval of the attribute a to the range V_a, where $a \in R$, and c represents the real number set. Provided that the conditions that $R = M \cup N$, $M \cap N = \emptyset$, and $N \neq \emptyset$ are met, then the information system S can also be called the decision system, in which the M and N are the condition attribute set and decision attribute set respectively.*

Definition 2.3. *In the information system $S =< U, R, V, f >$, $\exists \forall W \subseteq R$, where W is the subset of R, which contributed to the undivided relationship, the equivalence relationship $IND(W)$ (also known as U/W):*

$$IND(W) = \{(x, y) \in U \times U | \forall w \in W, f(x, w) = f(y, w)\} \tag{1}$$

Definition 2.4. *Supposing that $\{(a, c_1^a), (a, c_2^a), \ldots, (a, c_{ka}^a)\}$ is the set of the split intervals of attribute domain $V_a = [l_a, r_a]$, where $l_a = c_0^a < c_1^a < c_2^a < \cdots < c_{ka}^a < c_{ka+1}^a = r_a$, dividing the attribute a into $m_a + 1$ equivalence class, and the c_k^a is the split point. Making the domain equal $P_a = \left\{[c_0^a, c_1^a), [c_0^a, c_2^a), \ldots, [c_{ka}^a, c_{ka+1}^a]\right\}$, then, $P_a = \bigcup\limits_{a \in R} P_a$ is composed of a new decision information table:*

$$S^P =< U, R, V^P, f >, \quad f(x_a) = i \Leftrightarrow f(x_a) \in \left[c_i^a, c_{i+1}^a\right] \tag{2}$$

where $x_a \in U$, $i \in \{0, 1, \ldots, K_a\}$. The core essence of the discretization is to map the initial information system to a new interval information system.

3. **The Discretization Algorithm of Minimum Gini Index.** Gini index is used in CART algorithm, SLIQ algorithm, SPRINT algorithm when setting up the decision tree, the idea of which is calculating the impurity of each attribute and the Gini index, and then choosing the attribute with minimum Gini as the root node. Continuously, calculate the Gini index of other attribute until the final conditions are met: (1) the attribute chosen had shown in the former nodes or (2) the classes in the same interval are identical. The attribute with relatively small Gini index carries less class and more pure information. In this paper, we offer each column the same process to finish the division process, the theory illustration of discretization can be seen in Figure 1 and the concrete steps are as follows.

Step 1. Firstly, rank each attribute number $M_1, M_2, \ldots, M_n$, increasingly, then dealing with them by normalization making the number ranging from zero to one so that the same measure rule could be used. At the same time, the class set *ClassM* is also done with the step above. As soon as all the continuous numbers are set from 0 to 1, calculate the

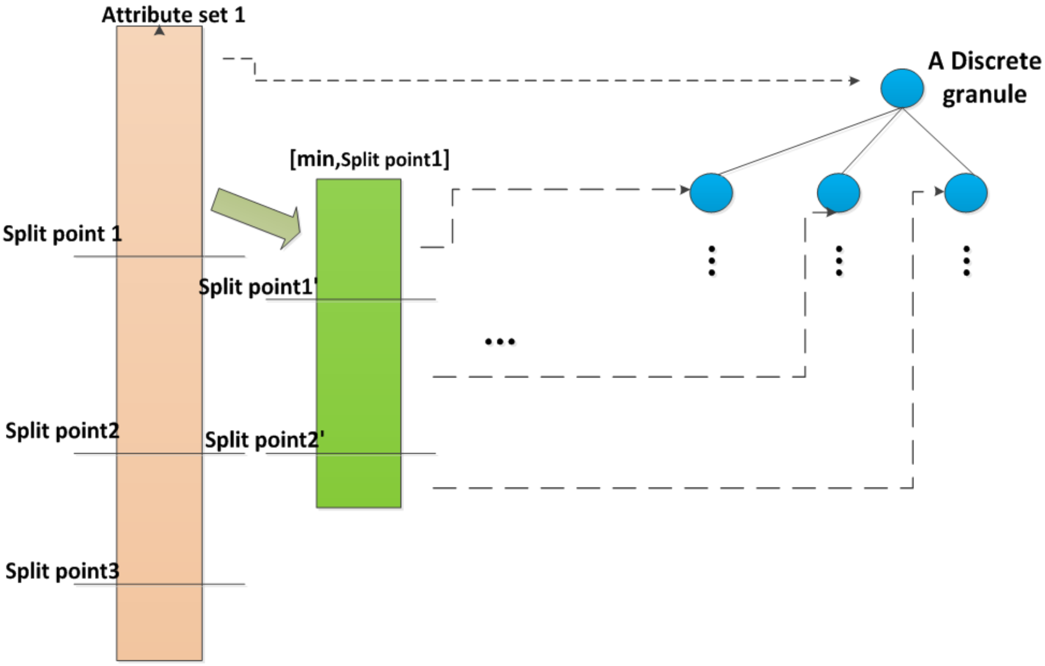

Figure 1. The theory illustration of discretization

potential split points – the midpoints of the neighboring number:

$$P = \frac{M_i + M_{i+1}}{2} \quad (i = 1, 2, \ldots, n-1) \tag{3}$$

Step 2. The potential split points $P_1, P_2, \ldots, P_{k-1}$ (k is the number of all objects) divide the set of attribute number into two ranges: $[\min, P)$ and $[P, \max]$. Then save the number of each class in $[\min, P)$ as $C1, C2, \ldots, Ct, \ldots, Cd$, where t is one of the classes. Calculate the impurity of the split point P:

$$Gini(P) = 1 - \sum_{i=1}^{d} p^2 = 1 - \left(\frac{C1}{d}\right)^2 - \cdots - \left(\frac{Ct}{d}\right)^2 - \cdots - \left(\frac{Cd}{d}\right)^2 \tag{4}$$

where d is the number of the classes in the range $[\min, P)$, and p is the ratio of Ct and d. The identical method could be used to calculate the *Gini* index in the interval $[P, \max]$. Supposing that the attribute set M is divided into N subsets S_j ($j = 1, 2, \ldots, N$), then the *Gini* index $Gini_{split}$ can be defined as:

$$Gini_{split}(M) = \sum_{j=1}^{N} \frac{s_j}{s} Gini(M_j) \tag{5}$$

where s_j is the number of samples in one certain class and s is the number of all classes. Besides, after comparing the index calculated above, the attribute with $\min(Gini_{split}(M))$ can be gotten as well.

Step 3. The steps discussed finished the first step of the discretization for one attribute, then saving the place of split point as w_i, and testing the purity from line $w_i + 1$ to the end of this column. Copying the objects that had been covered into a set context, and ending the discretization if all the continuous numbers are in the same class or all the objects are covered, otherwise, start over from Step 3 until the condition is met.

Step 4. Save the split points generated after covering all the objects in each attribute column. Represent the result by generating the formal context mapping the original data into the intervals.

4. **The Optimization of the Formal Context.** The coarse discretization process has been completed, and the formal context was generated. In order to get the better clustering effect and to get more detailed division, a new method for optimization to the rows and columns of the formal context is also proposed. The combination of these two methods could optimize the result. The essential steps are as follows.

Supposing that there is a formal context K consisting of m rows and n columns. Defining each attribute of each column as $a_{i1}, a_{i2}, \ldots, a_{ij}, \ldots, a_{im}$, then a new index *CGAO* (Combination of Gini And Objects) is introduced:

$$CGAO = Gini_{split(m_i)} + \frac{1}{\sum m_i} \tag{6}$$

where m_i is the attribute of i column. Then calculate the following formula:

$$Gini_{split}(m_1) + \frac{1}{\sum\limits_{i=1}^{m} a_{i1}}, Gini_{split}(m_2) + \frac{1}{\sum\limits_{i=1}^{m} a_{i2}}, \ldots, Gini_{split}(m_n) + \frac{1}{\sum\limits_{i=1}^{m} a_{in}} \tag{7}$$

Exchanging the column with minimum *CGAO* to the first column, then sorting the number in the first column in descending order, make the attribute number $a_{i1} = 1$ arranging continuously starting from a_{11}, that is to meet the condition that $\sum\limits_{i=1}^{q1} a_{i1} = q_1$ and $\sum\limits_{i=q_1+1}^{m} a_{i1} = 0$. Meanwhile, make the identical operation with the same new order to the class set. The formal context K is split into sub-matrix context $K1$ and $K2$. Then calculate the $CGAO$ from row q_1+1 to the ending row m:

$$Gini_{split}(m_2)+\frac{1}{\sum\limits_{i=q_1+1}^{m}a_{i2}},\ Gini_{split}(m_3)+\frac{1}{\sum\limits_{i=q_1+1}^{m}a_{i3}},\ \ldots,\ Gini_{split}(m_n)+\frac{1}{\sum\limits_{i=q_1+1}^{m}a_{in}} \quad (8)$$

Exchanging the column with minimum index calculated above with the second column, also, sorting the number in the second column in descending order, make the attribute number $a_{i2}=1$ arranging continuously starting from $a_{q_1+1,2}$, that is to meet the condition that $\sum\limits_{i=q_1+1}^{q_2}a_{i2}=q_2-q_1$ and $\sum\limits_{i=q_2+1}^{m}a_{i2}=0$. The first granule layer will be gotten until all the objects are covered. Do the same steps on each formal context gained from the steps above.

Based on the steps given above, the coarse granules are divided into more detailed granule layers, which completed the clustering process, making the granules in the same layer tight in the class and loose during the classes, which contributes to the knowledge discovery and other data mining supervised.

5. **Verification with Pattern Recognition System.** In order to illustrate the effect of discretization, a pattern recognition system that could process the mixed data was designed to test the result, and the illustration can be seen in Figure 2. The basic steps of a pattern recognition system are: data testing, preprocessing, feature extraction, classification and result representation, which are the key ways to test the result. We chose the random forest classifier as the tool. To illustrate the process discussed above, the iris database was chosen, which contains 150 objects, 4 attributes: sepal length, sepal width, petal length, pedal width (representing with *attr*1, *attr*2, *attr*3, *attr*4 respectively in Table 2), and one decision attribute concluding three classes: Setosa, Versicolor and VirGinica.

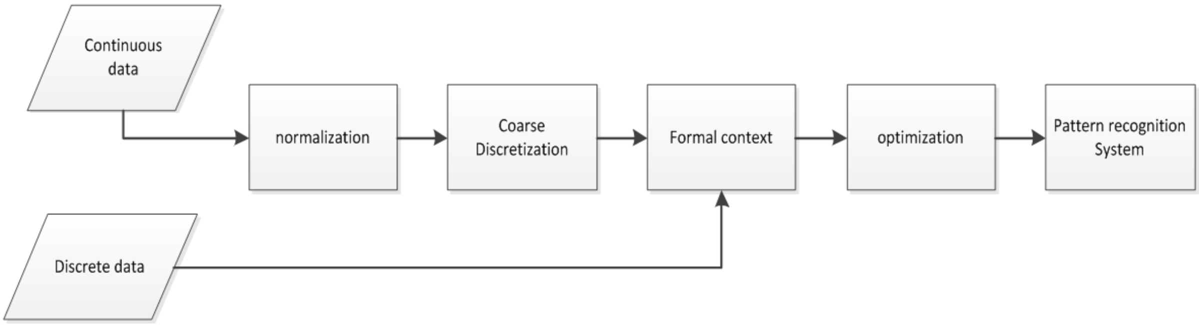

FIGURE 2. The flow chart of discretization

TABLE 2. Split points information of iris database

	Split point 1	*Split point* 2	*Split point* 3	*Split point* 4	*Number_points*
*attr*1	0.6859	0.7756	0.891	/	3
*attr*2	0.4167	0.4808	/	/	2
*attr*3	0.3013	0.5962	0.6346	0.6474	4
*attr*4	0.0897	0.2115	0.2244	/	3

With the calculation of the split points, the initial data was mapped to the sole interval, representing with number and letter. For instance, $A1$, $A2$, $A3$, $A4$ are used to represent the intervals $[0,0.6859]$, $(0.6859,0.7756]$, $(0.7756,0.8910]$, $(0.8910,1]$. $B1$, $B2$, $B3$ stand for intervals $[0,0.4167]$, $(0.4167,0.4808]$, $(0.4808,1]$, $C1$, $C2$, $C3$, $C4$, $C5$ represent the intervals $[0, 0.3013]$, $(0.3013, 0.5962]$, $(0.5962, 0.6346]$, $(0.6346, 0.6474]$, $(0.6474, 1]$, and $D1$, $D2$, $D3$, $D4$, map the intervals $[0, 0.0897]$, $(0.0897, 0.2115]$, $(0.2115, 0.2244]$, $(0.2244, 1]$. Iris database was integrated, which can be seen in Table 2. And the partial integrated iris data is in Table 3.

Table 3. Partial integrated iris data

Objects	*attr*1	*attr*2	*attr*3	*attr*4
1	$A1$	$B2$	$C1$	$D1$
2	$A1$	$B1$	$C1$	$D1$
3	$A1$	$B1$	$C1$	$D1$
...	...	...	...	...
15	$A2$	$B3$	$C1$	$D1$
...	...	...	...	...

Generate the initial formal context with the information, then optimize it with the method discussed above to improve the effect of clustering and the precision. The partial formal context optimized can be seen in Table 4. Basic clustering effect making the number “1” and “0” gathered is obvious.

Table 4. Partial formal context optimized

0	5	10	7	···	6
54	1	1	1	···	0
84	1	1	1	···	0
120	1	1	1	···	0
135	1	1	1	···	0
...	...	...	...	...	...
150	1	0	1	0	0
1	0	0	0	0	0
...	...	...	...	...	...

Other data from UCI database is used to illustrate the generalization ability of the method. The detailed information about the data can be seen in Table 5. The columns increased after coarse discretization, which tested the discretization effect dealing with relatively high dimensional data.

Table 5. Information of the five databases in UCI

Data set	*Features*	*Objects*	*Class*
Glass	9	214	6
Wine	14	178	3
Heart	13	270	2
Ecoli	7	336	2
Sonar	60	28	2
Iris	4	150	3
Ionosphere	35	351	2

Two other methods are used to compare the discretization effect. One is the equalization sub-box discretization, setting the number of intervals is 3, where the width depends on the number of the boxes, making the width of data in each box equal. The other is the equal frequency discretization that the number of each box is identical. As the widespread validation method, Leave One Out Cross Validation (LOOCV) is used to make the most of the sample data, the idea of which is that supposing that there are N samples, $N - 1$ samples will be used to train, and one of the left will be tested to evaluate the precision. Although the amount of calculation grows increasingly when the times grows, the method unbiased. The random forest classifier, integrating many CART classification and sampling data and choosing attributes randomly, is used to test the samples, which avoided the overfitting problem. The comparison result is shown in Table 6.

Table 6. The classification mean error rate

Data set	*Classification error rate*		
	Method 1 *(%)*	*Method* 2 *(%)*	***Our Method (%)***
Glass	46.729	30.373	23.545
Wine	7.303	4.4944	0.565
Heart	22.222	21.851	14.5
Ecoli	4.166	7.142	1.493
Sonar	19.711	21.153	0
Iris	4	5.3	2.12
Ionosphere	8.5	9.7	5.9

After analyzing Table 6, the effect of the method we proposed is better than the other two classical methods on precision, especially on the wine and sonar data set, the precision is 99.435% and 100%, besides, and the best precision of wine data set can achieve 100%.

6. **Conclusion.** In this paper, we proposed a discretization method based on minimum Gini index and optimization on formal context, which realized the function that dealt with continuous data in information system and extracted the formal context that can be used in FCA. Calculating the minimum Gini index of each attribute until covering all the objects to finish the coarse division, then the optimized split points are found. Exchange columns with minimum CGAO to start from the first column to complete the optimization to finish the detailed division, which is based on the supervision discretization. Finally, the result tested by the pattern recognition system designed showed that the method designed that ensured the precision laid the foundation to the Formal Concept Analysis on continuous data set. The method we proposed enriched the discretization methods, which will, to some extent, contribute to the data mining process and the application of Formal Concept Analysis theory.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61273019, 81373767, and 61501397. It is also supported by the Natural Science Foundation of Hebei Province under the Grant No. F2016203443. The authors gratefully acknowledge the supporters.

REFERENCES

[1] R. Wille, Restructuring lattice theory: An approach based on hierarchies of concepts, in *Ordered Sets*, I. Rival (ed.), Dordrecht-Boston, 1982.

[2] P. Jonas, I. I. Dmitry, O. K. Sergei and D. Guido, Formal concept analysis in knowledge processing: A survey on applications, *Expert Systems with Applications*, vol.40, no.16, pp.6538-6560, 2013.

[3] A. Simon and O. Constantinos, Discovering knowledge in data using formal concept analysis, *International Journal of Distributed Systems and Technologies*, vol.4, no.2, pp.31-50, 2013.

[4] M. Nida, K. Hela and M. Mondher, Parallel learning and classification for rules based on formal concepts, *Procedia Computer Science*, vol.35, pp.358-367, 2014.

[5] Z. Pei, D. Ruan, D. Meng and Z. Liu, Formal concept analysis based on the topology for attributes of a formal context, *Information Sciences*, vol.236, no.1, pp.66-82, 2013.

[6] H. Li, C. Mei and Y. Lv, Incomplete decision contexts: Approximate concept construction, rule acquisition and knowledge reduction, *International Journal of Approximate Reasoning*, vol.54, no.1, pp.149-165, 2013.

[7] E. Q. Abderrahim, A. Driss and E. Yassine, Formal concept analysis for information retrieval, *International Journal of Computer Science and Information Security*, vol.7, no.2, pp.117-121, 2010.

[8] M. F. Usama and K. B. Irani, Multi-interval discretization of continuous-valued attributes for clasafication leaning, *Proc. of the 13th International Joint Conference on Artificial Intelligence*, vol.2, pp.1022-1027, 1993.

[9] S. H. Nguyen, Some efficient algorithms for rough set methods, *Proc. of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Granada, Spain, pp.1451-1456, 1996.

[10] R. Susmaga, Analyzing discretizations of continuous attributes given a monotonic discrimination function, *Intelligent Data Analysis*, vol.1, no.3, pp.157-179, 1997.

[11] C. Chen, Z. Li, S. Qiao and S. Wen, Study on discretization in rough set based on genetic algorithm, *Proc. of the 2nd International Conference on Machine Learning and Cybernetics*, Xi'an, vol.3, pp.1430-1434, 2003.

[12] X. Li, A new method based on cloud model for discretization of continuous attributes in rough sets, *Pattern Recognition and Artificial Intelligence*, vol.16, no.1, pp.33-38, 2003.

[13] H. Xie, Discretization of continuous attributes in rough set theory based on information entropy, *Chinese Journal of Computers*, vol.28, no.9, pp.1570-1574, 2005.

[14] T. Zhang, H. Shi and Z. Li, Visual discretization for decision continuous formal context, *Application Research of Computers*, vol.33, no.2, pp.388-391, 395, 2016.