International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 11, No. 1 (2018), pp.23-36
http//dx.doi.org/10.14257/ijsip.2018.11.1.02

A Rough Set Theory based Approach for Massive Data Mining

Xianyang Li[1, 2], Guihua Qiu[1, 2]* and Anshan Lu[1, 2]

[1]*College of Electronics and Information Engineering, Qinzhou University, Guangxi, China*
[2]*Key Laboratory for Electronic Devices Inspection, Qinzhou, Guangxi, China*
5579934@qq.com

Abstract

As the wide use of digital devices and Internet-based applications, the amount of data on the whole world is growing significantly. For facing the massive of data sets like texts, videos, images, GPS, even medical information, data mining would be a suitable approach to make full use of the data for supporting advanced decision-makings. This paper introduces a rough set theory-based data mining approach for massive data sets. This approach uses an innovative discretization method to make the decision table to be compatible. Three steps with suitable models or algorithms are equipped to the clustering method. Based on the clustered data, the rough set-enabled data mining approach is introduced. Experiments are carried out through three levels to test the feasibility of the proposed approach by comparing the Greedy algorithm, information entropy-based algorithm, and importance-based algorithm. Several contributions are significant from this paper. Firstly, it could be observed that after using the proposed approach, the amount of breakpoints candidates has been greatly reduced. Some experiments achieve about 85% reduction. The reason is attributed to the Step 3 which makes the time complexity achieve $O(|U|\times|C|^2\times w^2\times\log|U|)$. *Secondly, it could be observed that the proposed approach outperforms the other three methods in the computational time. As the increasing of breakpoints candidates, the advantages of the proposed algorithm are more obvious. Finally, the information entropy-based approach (III) is prone to memory overflow. That means this approach takes much more memory to carry out the calculation.*

Keywords*: Data Mining, Rough Set Theory, Massive, Algorithm, Training*

1. Introduction

As the wide use of digital devices and Internet-based applications, the amount of data on the whole world is set to grow 10-fold in the next six years to 2020 from around 4.4 zettabytes to 44ZB [1]. According to IDC's annual Digital Universe study, it is predicted that, by 2020, the amount of information produced by machines, the so-called Internet of Things (IoT), will account for about 10% of data on earth [2]. Key predictions included: that by 2020, one tenth of the world's data will be produced by machines [3]; that the amount of useful data produced will increase from 22% in 2013 to more than 35% in 2020; most of the world's data will be produced in emerging markets [4]; the amount of data that spends some of its lifetime in the cloud will double; and the amount of data will increasingly outpace available storage [5]. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s. It has been suggested by Nick Couldry and Joseph Turow that practitioners in Media and Advertising approach as many actionable points of information about millions of individuals [6]. The industry appears to be moving away from the traditional approach of using specific media

Received (April 2, 2017), Review Result (November 24, 2017), Accepted (December 1, 2017)

ISSN: 2005-4254 IJSIP

environments such as newspapers, magazines, or television shows and instead tap into consumers with technologies that reach targeted people at optimal times in optimal locations.

For facing the massive of data sets like texts, videos, images, GPS, even medical information, data mining would be a suitable approach to make full use of the data for supporting advanced decision-makings. Data mining refers to an interdisciplinary subfield of computer science which is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [7]. The overall goal of the data mining process is to extract key information from a data set and transform it into an understandable structure for further use which may involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating [8]. Data mining is the analysis step of the "knowledge discovery in databases" (KDD) process.

However, when using the data mining to deal with huge number of data, it is sometimes difficult to determine the exact the values. Thus, rough set theory could be used in such cases. Rough set is a formal approximation of a crisp set (*i.e.*, conventional set) in terms of a pair of sets which gives lower and the upper approximation of the original set [9]. The idea of rough set was proposed by Pawlak (1981) as a new mathematical tool to deal with vague concepts [10]. Since the development of rough sets, extensions and generalizations have continued to evolve with the initial developments focusing on the relationship - both similarities and difference - with fuzzy sets. While some literature contends these concepts are different, other literature considers that rough sets are a generalization of fuzzy sets - as represented through either fuzzy rough sets or rough fuzzy sets [11, 12]. Rough set methods can be applied as a component of hybrid solutions in machine learning and data mining. They have been found to be particularly useful for rule induction and feature selection (semantics-preserving dimensionality reduction). Rough set-based data analysis methods have been successfully applied in bioinformatics, economics and finance, medicine, multimedia, web and text mining, signal and image processing, software engineering, robotics, and engineering (*e.g.* power systems and control engineering) [13]. Recently the three regions of rough sets are interpreted as regions of acceptance, rejection and deferment, which leads to three-way decision making approach with the model which can potentially lead to interesting future applications.

There are some challenges when using the rough set theory for mining some massive data sets. Firstly, the decision table may greatly influence the mining results. The discretization consequences from decision tables directly influence the identification ratio of the samples [14]. Secondly, the rough set-enabled data mining is based on equal relationship or non-identical relations, while the decision table may be intended to discretize data such as calculation of property core, property simplicity, and value simplicity. Thirdly, facing the massive data sets, the performance of algorithms or models is low. The balance of high efficiency and high identification ratio is challenging.

In order to deal with the above challenges, this paper introduces a rough set theory-based data mining approach for massive data sets. This approach uses an innovative discretization method to make the decision table to be compatible. Three steps with suitable models or algorithms are equipped to the clustering method. Based on the clustered data, the rough set-enabled data mining approach is introduced. Experiments are carried out through three levels to test the feasibility of the proposed approach by comparing the Greedy algorithm, information entropy-based algorithm, and importance-based algorithm. In order to present the paper logically and smoothly, this paper is organized as follows: Section 2 gives a briefly discussion about the discretization algorithm for proposed approach. Section 3 presents the proposed clustered algorithm.

Section 4 illustrates the rough set-based data mining approach. Section 5 gives the experiments and discussions. Conclusions are given in Section 6 to finalize this paper by giving the future work.

2. Discretization Algorithm for Proposed Approach

The discretization algorithm for keeping the compatibility of the decision table includes three major steps: 1) calculate the breakpoint candidate set for each property from the decision table; 2) within the obtained candidate set of breakpoints, heuristic algorithm is used for picking the breakpoints; 3) the selected breakpoints are used for the discretization of the decision table.

2.1. Step 1

Assume that a decision table $S =< U, A = C \cup D, V, f >$, for each property in S, the following operations will be carried out:

(1) If the element property is varchar, based on the categories of the varchar, each object of strings is converted into integer whose value is used for sequencing.

(2) If the element property is Boolean type, these properties could be converted into '0' and '1' and then sequenced.

(3) If the element property is integer or float, they are sequenced by their values.

(4) Assume a property $a(a \in C)$, after sequencing, we can get:

$$l_a = v_0^a < v_1^a < v_2^a < ...v_{n_a}^a = r_a \tag{1}$$

The breakpoint candidates could be selected using

$$c_i^a = (v_{i-1}^a + v_i^a)/2 \quad i = 1, 2, ..., n_a \tag{2}$$

Using this method, the compatibility of decision table could be ensured. This method also has high efficiency given its time complexity is $O(|C| \times |U| \times \log|U|)$.

2.2. Step 2

For the second step, some heuristics algorithms could be used. Assume the average amount of the conditional property is w. The targeted number of selected breakpoints is N. The average amount of breakpoints is $w \times |C|$. For different heuristics algorithms, the time complexity would be different as follows: Greedy algorithm ($O(|U|^2 \times |C| \times w \times N)$), information entropy-based discretization algorithm ($O(|U| \times |C|^2 \times w^2 \times \log|U|)$), and property importance-based method ($O(|U| \times |C| \times w \times \log|U|)$).

Besides the time complexity, the above mentioned algorithms have some cons and pros. For the greedy algorithm, the breakpoints candidates will be influenced by each other so that the targeted points will be better than the other two. But the time complexity is higher. For the information-based discretization algorithm, the calculation for the rest of the points will be carried out after selecting one point [16]. That makes the time complexity is higher, while the obtained results are good. For the property importance-based method, stepwise deletion concept is used so that the

time complexity is lower. However, the obtained breakpoints will be non-averagely distributed. That results in low identification ratio.

2.3. Step 3

For the step 3, the data from original decision table could be numbered by 0,1,2, on the properties according to the selected breakpoints. Thus, the discretization of decision table could be finished. The time complexity is $O(|U|\times|C|)$.

From the above analysis, the second step will greatly influence the discretization in terms of efficiency and discrete results. Two considerations should be taken. First, in order to improve the algorithm speed, it is better to reduce the targeted breakpoints. Second, the influence degrees among different breakpoints could be used for the candidate selection so that the amount of discrete breakpoints could be reduced. Thus, if the candidate breakpoints could be clustered given a single property, the amount could be reduced [17]. After that, for all properties, the selections could be carried out which may implement the discretization of decision tables efficiently and effectively.

3. Proposed Clustering Approach

Based on the discussion in previous sections, an efficient data mining approach is necessary for the decision table discretization, for example a cluster method. For most of the data from real-life cases, statistics trends could be applied [18-20]. Thus, the breakpoints candidates should follow a certain distribution. In this section, we propose a rough set cluster algorithm given the importance of the breakpoints.

The importance of a breakpoint is determined by the quantity of identified entities [21]. For a breakpoint c_i^a, the quantity of identified entity is $W^X(c_i^a)$, where c_i^a is the no. i breakpoint of property a, $1\le i\le n_a$. n_a is the total number of breakpoints of property a. U is the entity set and $X\subseteq U$ is the set obtained from separating from c_i^a. The decision attribute value is $j=1,2,...,r$. r is the types of decision. The entity quantity could be calculated by:

$$l_j^X(c_i^a)=|\{x\,|\,x\in X\wedge(a(x)<c_i^a)\wedge(d(x)=j)\}| \tag{3}$$

For the decision attribute value j, the entity quantity is

$$r_j^X(c_i^a)=|\{x\,|\,x\in X\wedge(a(x)\ge c_i^a)\wedge(d(x)=j)\}| \tag{4}$$

The above (3) and (4) present the property value of a is less and greater-or-equal than the value of breakpoint c_i^a respectively. Thus, we can get

$$l^X(c_i^a)=\sum_{j=1}^{r}l_j^X(c_i^a)=|\{x\,|\,x\in X\wedge(a(x)<c_i^a)\}| \tag{5}$$

$$r^X(c_i^a)=\sum_{j=1}^{r}r_j^X(c_i^a)=|\{x\,|\,x\in X\wedge(a(x)\ge c_i^a)\}| \tag{6}$$

From (5) and (6), we can get

$$W_X(c_i^a) = l^X(c_i^a) \times r^X(c_i^a) - \sum_{i=1}^{r} l_i^X(c_i^a) \times r_i^X(c_i^a) \tag{7}$$

The importance $W_X(c_i^a)$ reflects the relation between conditional property value and decision property value. The value of $W_X(c_i^a)$ is bigger, the more important the breakpoint c_i^a is. That point could be selected to build up the breakpoint set. For the clustering process, based on the importance, single property clustering could be achieved. During the clustering, stepwise trial method is used for each property [22]. Therefore, the threshold should be determined. The threshold e could be calculated by:

$$e = \sqrt{\frac{\sigma^2}{n}} = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{n} \tag{8}$$

Where, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, n is the training samples amount.

Based on the concept, a dynamic data mining (clustering) algorithm is proposed as follows:

Input: Decision table $S =< U, A = C \cup D, V, f >$ and breakpoint candidate set $CCut$

Output: The first sorted breakpoint set CUT_f from S

Step 1: Calculate the importance of each breakpoint in $CCut$

for $i = 1$ to $|C|$ do

According to the property c_i, calculate the importance of each breakpoint given their sequences;

endfor

Step 2: for $i = 1$ to $|C|$ do

Calculate the maximum value of importance in c_i : $IMP^M = \max(W^X(c_j^{c_i}))(j = 1, 2, ..., n_{c_i})$;

Normalize the importance and keep in $IMP[][]$, $IMP[i][j] = W^X(c_j^{c_i}) / IMP^M$;

According to the peak value of breakpoint importance, divide them into two parts: $Left_i$ and $Right_i$;

$Left_i = Right_i = \varnothing$; $k = 1$;

While $(IMP[i][k] < 1)$ Do $Left_i = Left_i \cup \{c_k^{c_i}\}$; $k = k + 1$;Loop

While $k \le n_{c_i}$ Do $Right_i = Right_i \cup \{c_k^{c_i}\}$; $k = k+1$;Loop

Step 3: Cluster the $Left_i$;

Calculate the dynamic threshold $e = \frac{\sqrt{\sum_{j=1}^{Left_i} (IMP[i][j] - \bar{x})^2}}{|Left_i|}$

$(1 \le j \le |Left_i|, \bar{x} = (\sum_{j=1}^{|Left_i|} IMP[i][j])/|Left_i|)$

Initial the cluster category $K = 1$, the iteration variable $v = e+1$;

While $(v > e)$ Do

Establish K initial cluster center, select K breakpoint from $Left_i$, their importance is regarded as the center;

Set the loop controller $e_1 = 0$;

While $(e_1 \ne v)$ Do

$e_1 = v$;

Calculate the importance of each breakpoints in $Left_i$ and their distance to K center. Modify the center according to the average value. Recalculate the standard deviation s_i.

Assume that the distribution of each category U_i is $y_1, y_2, ..., y_{|U_i|}$, $\bar{y} = (\sum_{j=1}^{|U_j|} y_j)/|U_i|$;

$s_i = \sqrt{\frac{\sum_{j=1}^{|U_j|} (y_j - \bar{y})^2}{|U_i|}}$, $v = (\sum_{i=1}^{K} s_i)/K$;

Loop

$K = K+1$;

Loop

Select the minimum value of breakpoint in each category and then add into CUT_f;

Do the same operations on the $Right_i$;

endfor

Return CUT_f;

From the above description of the proposed algorithm, the time complexity is critical. Assume the average number of condition property and breakpoints are w and $w\times|C|$ respectively, the complexity for step 1 is $O(w\times|C|\times|U|\times\log|U|)$. The time complexity for step 2 is $O(k^2\times|C|\times|U|)$. Then the total time complexity and space complexity are $O(|C|\times|U|\times(w\times\log|U|+k^2))$ and $O(w\times|C|+|U|)$ respectively.

4. Rough Set-based Data Mining Approach

After processing by the clustering method proposed in section 3, the clustered breakpoints have much reduced quantity. However, they are not suitable for the final decision tables [23-24]. First of all, the clustered data only considers single condition property not taking account the whole property set. Secondly, for more properties, the clustered data will have many redundancy. Thus, a data mining approach should be used for further extract the key information which will be presented by rough set theory.

Assume the data $X_1, X_2, ..., X_m$ come from the breakpoints which are dived by set P, the entity quantity identified by $c \notin P$ is:

$$W_p(c) = W^{X_1}(c) + W^{X_2}(c) + ... + W^{X_m}(c) \tag{9}$$

Where $W^{X_i}(c)$ is the data from X_i which could be identified by c.

Based on the definition, the data mining approach could be described as follows.

Input: Decision table $S = < U, A = C \cup D, V, f >$

Output: CUT^F and S^F

Step 1: Calculate the breakpoint set $CCut$ from S

for $i = 1$ to $|C|$ do

$CCut_i = \varnothing$;

In the property c_i, sequence all the entity value increasely;

If the value of c_i is non-integer or non-float, 0, 1, 2… will be used for presenting the value;

Calculate the rough set value for property c_i $c_j^{c_i} = (v_{j-1}^{c_i} + v_j^{c_i})/2$, $j = 1, 2, ..., n_{c_i}$;

$CCut_i = CCut_i \cup \{c_j^{c_i}\}$;

endfor

Step 2: Calculate the positive area $POS_C(D)$ of S;

Calculate the CUT^F. Use the CUT^F to do the discretization of S, the discrete decision table is S_1;

Calculate the positive area $POS'_C(D)$ of S_1;

If $|POS_C(D)| \neq |POS'_C(D)|$, then go step 3;

Else $\forall_{x \in X}(x \in POS_C(D)) \Rightarrow \forall_{y \in X}$;

Step 3: Let L is the set divided by CUT^F, $CUT^F = \varnothing$, $L = \{U\}$

For each $c \in CUT^F$, calculate $W_{CUT^F}(c)$;

Select the maximum breakpoint $c_{\max}$ from $W_{CUT^F}(c)$ and add into CUT^F;

$CUT^F = CUT^F \cup \{c_{\max}\}$;

For all $X \in L$, if $c_{\max}$ divide X into X_1 and X_2 rough sets, then delete the X from L. Add X_1 and X_2 into L;

Use the CUT^F to discretize S, get the new S^F;

Return CUT^F and S^F

Assume the average number of condition property and breakpoints are w and $w \times |C|$ respectively, the time complexity for step 1 is $O(|U| \times |C| \times \log|U|)$. Step 2 has the time complexity $O(|U| \times (|C| + \log|U|))$. And step 3 has the time complexity $O(|U| \times |C| \times (w \times \log|U| + k^2))$. The total time and space complexity of the algorithm are $O(|U| \times |C|^2 \times w^2 \times \log|U|)$ and $O(w \times |C| + |U|)$.

5. Experiments and Discussions

In order to evaluate the feasibility of the proposed algorithm, three tests are carried out. In these tests, we compared with greedy algorithm (I), importance-based algorithm (II), information entropy-based approach (III) and the proposed approach (IV). Firstly, we selected the data set from UCI for experiment 1. Secondly, Must Clean2 testing data set is used for experiment 2. And finally, Poker-Hand set is used for experiment 3. The total volume of the data set for each experiment is over 1 T.

5.1. Experiment 1

This experiment aims to testing the efficiency and effectiveness of the proposed approach. 7 groups of data from UCI database are used of this test [25]. Table 1 presents the experiment data samples ($\times 10^{10}$) including record amount and condition property amount. For each group, half of the data is randomly selected for training while the rest is for testing.

Table 1. UCI Data Samples

Sample Data	Record Amount	Condition Property Amount	Decision Category
Iris	154	4	2
Wine	187	13	3
Glass	241	8	3
Ecoli	363	7	4
Breast	694	9	3
Pima	786	8	3
Letter	21035	18	2

Table 2 presents the experimental results including the obtained breakpoints, accuracy rate, and time cost (Unit of Time).

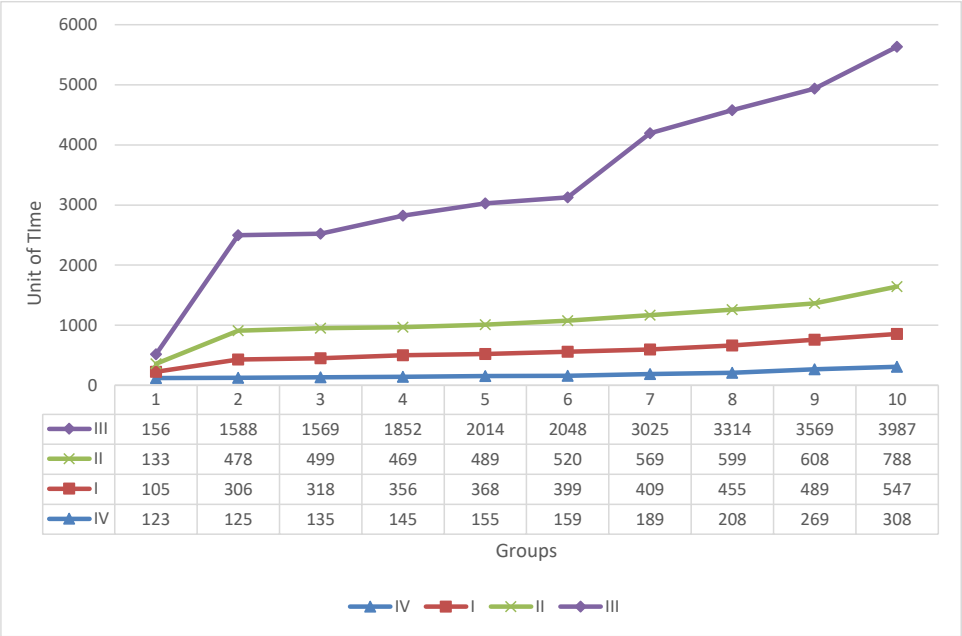
Table 2. Experiment Results

Data / Items	Iris			Wine			Glass			Ecoli			Breast			Pima			Letter		
Initial B	119			1263			921			356			80			1246			237		
Clustered	29			150			102			53			26			108			106		
I	6	92%	1.01	6	93.3%	22.1	11	63.9%	42.2	12	61.8%	32.03	11	94.6%	11	17	68.6%	680.3	63	74.2%	6524.6
II	5	92%	0.22	6	88.8%	1.8	11	63.4%	5.82	13	58.6%	4.79	12	94.3%	0.76	17	68.5%	49	62	77.5%	250.8
III	5	92%	0.17	5	90.2%	1.4	11	63.4%	7.68	12	60.4%	5.81	11	94.3%	0.82	18	67.2%	51.5	81	77.2%	268.6
IV	6	92%	0.11	6	94.4%	0.69	12	69.3%	1.18	14	65.7%	1.53	13	95.3%	0.56	20	69.1%	6.34	89	80.2%	114.1

From Table 2, it could be observed that after using the proposed approach, the amount of breakpoints candidates has been greatly reduced. Some experiments achieve about 85% reduction such as set Wine. Compared with the three other methods, the proposed approach has better accuracy rate for each data group. The processing time spent is significantly reduced so that more data could be processed. The reason is attributed to the Step 3 which makes the time complexity achieve $O(|U|\times|C|^2\times w^2\times\log|U|)$. The dynamic and rough set value are combined to consider the data distribution so that the data sets are always compatible. Therefore, the total time complexity of the proposed approach could be calculated as $O(|U|\times|C|\times(w\times\log|U|+k^2))+O(|U|\times|C|^2\times k^2\times\log|U|)$. That is why the proposed approach uses much less time on the computational works.

5.2. Experiment 2

In order to test the proposed approach on large amount of breakpoints candidates, Musk Clean 2 data sets from UCI are used for the second experiments. We selected 659,618,022,293,006 pieces of records as data samples. 10 groups of breakpoint candidate are set as 29,399; 38,256; 43,022; 47,006; 49,064; 51,077; 52,661; 54,000; 55,137; and 56,129. Figure 1 presents the experimental results from four methods.



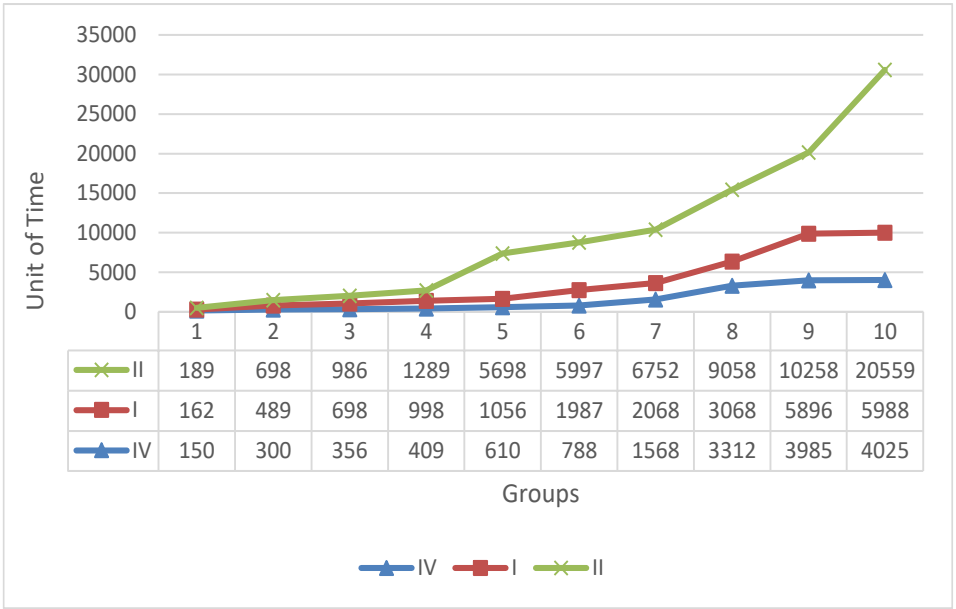
	1	2	3	4	5	6	7	8	9	10
III	156	1588	1569	1852	2014	2048	3025	3314	3569	3987
II	133	478	499	469	489	520	569	599	608	788
I	105	306	318	356	368	399	409	455	489	547
IV	123	125	135	145	155	159	189	208	269	308

Figure 1. Experimental Results -2

From Figure 1, it could be observed that the proposed approach outperforms the other three methods in the computational time. As the increasing of breakpoints candidates, the advantages of the proposed algorithm are more obvious. And information entropy-based approach (III) is prone to memory overflow. That means this approach takes much more memory to carry out the calculation.

5.3. Experiment 3

Experiment 3 aims to test the efficiency and effectiveness of various approaches on massive data sets. We selected Poker-Hand data from UCI for this experiments. This data set volume has over 10 T where we use 10 condition properties for the testing. Figure 2 shows the experiment results.



	1	2	3	4	5	6	7	8	9	10
II	189	698	986	1289	5698	5997	6752	9058	10258	20559
I	162	489	698	998	1056	1987	2068	3068	5896	5988
IV	150	300	356	409	610	788	1568	3312	3985	4025

Figure 2. Experiment Results - 3

From Figure 2, the proposed approach outperforms the other two: greedy algorithm (I) and importance-based algorithm (II). The information entropy-based approach (III) doesn't have the results because this approach heavily depends on the memory size. In this experiment, the memory overflow is occurred. It could be observed that, the proposed approach has better computational time which is about 25% saving of time.

From the above three experiments, the proposed approach is able to efficiently and effectively perform the clustering and mining purposes. From the computational cost perspective, this approach achieves better performance. It can save the computational time about 25% averagely. In some cases, the computational saving can reach approximately 80% compared with other approaches mentioned in this paper like information entropy-based approach which accounted memory overflow in experiment 3.

6. Summary

This paper talks about a data mining approach for massive data sets using rough set theory. This approach uses an innovative discretization method to make the decision table to be compatible. Three steps with suitable models or algorithms are equipped to the clustering method. Based on the clustered data, the rough set-enabled data mining approach is introduced. Experiments are carried out through three levels to test the feasibility of the proposed approach by comparing the Greedy algorithm, information entropy-based algorithm, and importance-based algorithm.

Several findings are significant from this paper. Firstly, it could be observed that after using the proposed approach, the amount of breakpoints candidates has been greatly reduced. Some experiments achieve about 85% reduction. The reason is attributed to the Step 3 which makes the time complexity achieve $O(|U|\times|C|^2\times w^2\times\log|U|)$. Secondly, it could be observed that the proposed approach outperforms the other three methods in the computational time. As the increasing of breakpoints candidates, the advantages of the proposed algorithm are more obvious. Finally, the information entropy-based approach (III) is prone to memory overflow. That means this approach takes much more memory to carry out the calculation.

Future improvements could be carried out from two aspects. First of all, the lower approximation of a target set is a conservative approximation consisting of only those objects which can positively be identified as members of the set. In order to improve the rough set presented the data, the more precise or reasonable presentation could be worked out. Secondly, more experiments which are going to compare with more approaches used other parameters should be further investigated. The comparisons from the experiments could be used for working out more efficient and effective methodology for dealing with massive data sets which are more complex and abstract.

Acknowledgement

Authors would like to acknowledge the support from Research on the Basic Skills Improvement for Young Teachers in Guangxi University in 2017, Short Texts Research on the Construction Technology of the Mass Organizations of Media Information Structures (No.2017KY0795); Qinzhou Science and Technology Project: Development of IoT-based Aquatic Environment Monitoring System Key Technology Research (No. 20164410); Qinzhou Electronic Product Testing Key Laboratory Open-project Grant.

References

[1] G. Wang, A. Gunasekaran, E. W. Ngai and T. Papadopoulos, "Big data analytics in logistics and supply chain management: Certain investigations for research and applications", International Journal of Production Economics, vol. 176, **(2016)** pp. 98-110.

[2] R. Y. Zhong, G. Q. Huang, S. L. Lan, Q. Y. Dai, C. Xu and T. Zhang, "A Big Data Approach for Logistics Trajectory Discovery from RFID-enabled Production Data", International Journal of Production Economics, vol. 165, **(2015)**, pp. 260-272.
[3] M. Salehan and D. J. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics", Decision Support Systems, vol. 81, **(2016)**, pp. 30-40.
[4] R. Y. Zhong, Q. Y. Dai, T. Qu, G. J. Hu and G. Q. Huang, "RFID-enabled Real-time Manufacturing Execution System for Mass-customization Production", Robotics and Computer-Integrated Manufacturing, vol. 29, no. 2, **(2013)**, pp.283-292.
[5] L. Y. Pang, R. Y. Zhong, J. Fang and G. Q. Huang, "Data-source interoperability service for heterogeneous information integration in ubiquitous enterprises", Advanced Engineering Informatics, vol. 29, **(2015)**, pp. 549-561.
[6] S. Lan, H. Zhang, R. Y. Zhong, G. Huang and H. K. Chan, "A customer satisfaction evaluation model for logistics services using fuzzy analytic hierarchy process", Industrial Management & Data Systems, vol. 116, **(2016)**, pp. 1024-1042.
[7] R. Y. Zhong, G. Q. Huang, Q. Y. Dai and T. Zhang, "Mining SOTs and Dispatching Rules from RFID-enabled Real-time Shopfloor Production Data", Journal of Intelligent Manufacturing, vol. 25, **(2014)**, pp. 825-843.
[8] R. Y. Zhong, S. T. Newman, G. Q. Huang and S. L. Lan, "Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives", Computers & Industrial Engineering, vol. 101, **(2016)**, pp. 572-591.
[9] S. Das and A. Roy, "Signature Verification Using Rough Set Theory Based Feature Selection", in Computational Intelligence in Data Mining—Volume 2, ed: Springer, **(2016)**, pp. 153-161.
[10] E. J. Gardiner and V. J. Gillet, "Perspectives on Knowledge Discovery Algorithms Recently Introduced in Chemoinformatics: Rough Set Theory, Association Rule Mining, Emerging Patterns, and Formal Concept Analysis", Journal of chemical information and modeling, vol. 55, **(2015)**, pp. 1781-1803.
[11] J. Zhan, Q. Liu and B. Davvaz, "A new rough set theory: rough soft hemirings", Journal of Intelligent and Fuzzy Systems, vol. 28, **(2015)**, pp. 1687-1697.
[12] R. Y. Zhong, G. Q. Huang, S. Lan, Q. Dai, T. Zhang and C. Xu, "A two-level advanced production planning and scheduling model for RFID-enabled ubiquitous manufacturing", Advanced Engineering Informatics, vol. 29, **(2015)**, pp. 799-812.
[13] G. Cattaneo, G. Chiaselotti, D. Ciucci and T. Gentile, "On the connection of hypergraph theory with formal concept analysis and rough set theory", Information Sciences, vol. 330, **(2016)**, pp. 342-357.
[14] X. Qiu, H. Luo, G. Xu, R. Zhong and G. Q. Huang, "Physical assets and service sharing for IoT-enabled Supply Hub in Industrial Park (SHIP)", International Journal of Production Economics, vol. 159, **(2015)**, pp. 4-15.
[15] R. Y. Zhong, G. Q. Huang and Q. Y. Dai, "A Big Data Cleansing Approach for n-dimensional RFID-Cuboids", Proceeding of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2014), 21-23 May, Taiwan, **(2014)**, pp. 289-294.
[16] M. Hong, M. Razaviyayn, Z.-Q. Luo and J.-S. Pang, "A Unified Algorithmic Framework for Block-Structured Optimization Involving Big Data: With applications in machine learning and signal processing", Signal Processing Magazine, IEEE, vol. 33, **(2016)**, pp. 57-77.
[17] R. Y. Zhong, S. Lan, C. Xu, Q. Dai and G. Q. Huang, "Visualization of RFID-enabled shopfloor logistics Big Data in Cloud Manufacturing", The International Journal of Advanced Manufacturing Technology, vol. 84, **(2016)**, pp. 5-16.
[18] M. Majdi, S. Abdullah and N. S. Jaddi, "Fuzzy Population-Based Meta-Heuristic Approaches for Attribute Reduction in Rough Set Theory, World Academy of Science, Engineering and Technology", International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 9, **(2015)**, pp. 2065-2073.
[19] R. Y. Zhong, G. Q. Huang, Q. Y. Dai and T. Zhang, "Mining SOTs and Dispatching Rules from RFID-enabled Real-time Shopfloor Production Data", Journal of Intelligent Manufacturing, vol. 25, **(2014)**, pp. 825-843.
[20] H. Liu, H. Guo and C.-a. Wu, "Hyperbox Granular Computing Based on Distance Measure", International Journal of Control and Automation, vol. 9, **(2016)**, pp. 1-10.
[21] R. Y. Zhong, X. Xu, E. Klotz and S. T. Newman, "Intelligent Manufacturing in the Context of Industry 4.0: A Review", Frontiers of Mechanical Engineering, vol. 3, **(2017)**, pp. 616-630.
[22] R. Y. Zhong, Z. Li, A. L. Y. Pang, Y. Pan, T. Qu and G. Q. Huang, "RFID-enabled Real-time Advanced Planning and Scheduling Shell for Production Decision-making", International Journal of Computer Integrated Manufacturing, vol. 26, **(2013)**, pp. 649-662.
[23] P. Ashok and G. K. Nawaz, "Upgraded Rough Clustering and Outlier Detection Method on Yeast Dataset by Entropy Rough K-Means Method", World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 9, **(2016)**, pp. 1732-1739.
[24] R. Y. Zhong, C. Xu, C. Chen and G. Q. Huang, "Big Data Analytics for Physical Internet-based intelligent manufacturing shop floors", International Journal of Production Research, vol.55, **(2017)**, pp. 2610-2621.

[25] G. Yang and K. Chen, "A Scheme Design and Configuration Model of Mechanical Equipment Processing Technology based on Polychromatic Sets Theory", International Journal of Control and Automation, vol. 7, **(2014)**, pp. 137-144.

Authors

Xianyang Li, he received bachelor's degree in Computer Science and Technology from Gannan Normal University in 2002. From September 2002 to July 2014, he worked as lecturer in Ganzhou Teachers College China. From September 2014 to now, he worked as associate professor in Qinzhou University, Guangxi, China. His research interests include Computer Networks, Data Mining, and Education Technology. He has published several papers in international journals and conferences.

Guihua Qiu, he Professor Qiu is an associate professor in Qinzhou University, Guangxi, China. His research areas include Computer Education, Big Data, and Network Security.

Anshan Lu, he is a professor in Qinzhou University, Guangxi, China. His research areas include Electronic Technology, Chaos Control and Synchronization Technologies.

