sciendo

POLISH MARITIME RESEARCH Special Issue 2018 S3 (99) 2018 Vol. 25; pp. 136-140
10.2478/pomr-2018-0122

RESEARCH AND APPLICATION OF RULE UPDATING MINING ALGORITHM FOR MARINE WATER QUALITY MONITORING DATA

Qiuhong Sun[1,2]
Junhai Zhang[1]
Xinhang Xu[3]
[1] Postdoctoral research station of Geography, Hebei Normal University, Shijiazhuang, China
[2] Hebei University of Science and Technology, Shijiazhuang, China
[3] State Grid Hebei Electric Power Research Institute, Shijiazhuang, China

ABSTRACT

This paper studies the characteristics of marine water quality monitoring data monitored by photoelectric sensor network, mines the potential information from the massive data. on account of the continuous accumulation of monitoring data, this paper focuses on the study of database with numerical attribute and proposes a rule updating algorithm for solving the rule maintenance issues caused by changes in the database. according to the rule, the algorithm forms a new database from part of the original data and the new data, and searches the new database by random search, thus can avoid creating a large number of redundant rules and can quickly mine effective rules at the same time. experimental results show that this method not only can avoid mining in the whole original massive data, but also can improve work efficiency, and can quickly and effectively find new data and find useful rules in the data with high practicability.

Keywords: Marine water quality monitoring, database updating, data mining, rule updating

INTRODUCTION

At present, the monitoring data of marine water quality are mainly based on the static database for rule mining and output. In the actual water quality monitoring process, the server side will accumulate a large amount of data over time, and these data on the monitoring of water quality are constantly updating and changing. Therefore, in this case, it is a very important issue to update the association rules in the process of dynamic database changes, and that is a mining problem of incremental association rules. The traditional update algorithm uses Apriori algorithm, but this algorithm is inefficient due to frequent I/O access and it's hard to measure the support settings. If the setting is too low, a lot of redundant rules will be generated. If the setting is too high, some useful but low-supportive information will be filtered out [1–3]. Therefore, this paper studies a rule updating mining algorithm and applies this method to update the rules of ocean water quality monitoring data mining.

RULE UPDATING

The traditional work of association rules data mining is based on the static database and according to the pre-set minimum support degree and minimum credibility [4–6], while the update issue of the dynamic database mainly studies the data changes in the database, the changes in the pre-set minimum support degree or the minimum credibility, and how to update the association rules without overall re-mining on the basis of existing rules.

The current rule updating algorithm mainly aims at Boolean database mining, which can effectively solve the data update mining problems, such as D. W. Cheung et al.'s FUP and FUP2 algorithms, and Feng Yu et al.'s IUA algorithm [7–9]. All of these algorithms can make full use of the rules and information of the previous mining and conduct rule updating, which can avoid duplication work and improve work efficiency. However, in most actual production processes, the attribute values of various fields to be mined are not only Boolean type, and the numerical databases are more widely used [10–12]. In such a database, it is also necessary to update the rules, previous scholars have proposed to transform the numerical attribute database into Boolean, and conduct association rules mining. Since the water quality monitoring database belongs to the numerical database type, this paper mainly studies the rule updating problem of the numerical attribute database.

Rule updating refers to that, when the setting of the minimum support degree and the minimum confidence is unchanged, there are new data adding to the original database. Here marks the original database as DB, the new dataset is db, and we mainly studies how to put the original data and the new data set into a new database.

For the update of association rules, we start mainly from the following two situations [13–15]:

(1) When the minimum support degree and the minimum confidence are set, the database changes. How to generate new association rules when new data is added to the original database, how to delete some data from the original database and how to generate new association rules [16].

(2) How to generate new association rules when the data in the database is not changed, and the minimum support degree and the minimum confidence are changed [17].

As the water quality data monitored by the photoelectric sensor network continuously accumulates over time, the data in the database is getting more and more, and we mainly study the first updating situation above [18].

Add db into DB, for any set of items X, there are the following possibilities.

(1) X are frequent itemsets both in DB and db;

(2) X is a frequent itemset only in DB, but is not in db;

(3) X is a non-frequent itemset only in DB, and also is in db;

(4) X is not a frequent itemset either in DB or db.

RESEARCH ON THE RULE UPDATING ALGORITHM

RULES UPDATING ALGORITHM PRINCIPLE

During the monitoring process of the photoelectric sensor network, the water quality data is affected by factors such as climate, ocean currents and emergencies, the water quality monitoring data will change, the monitoring data will accumulate more and more over time, and the initial database will change constantly. The changing data will cause corresponding changes in the association rules, and we use the difference to reflect these changes. When the difference becomes larger, namely there are more and more data, new data may imply some new potential information [19]. Therefore, the algorithm in this paper first calculates the difference between DB and db, and then draws part of the data from the DB according to the degree of difference, forms a new database with db, and then carries out the mining operation based on adaptive immune genetic association rules. Such mining results can both retain some rules of high support degree in DB, and reflect the potential rules of new database, meanwhile shorten the mining time and improve the mining efficiency.

RULE UPDATING ALGORITHM DEFINITION

The proposed algorithm in the initial stage needs to be calculated by the following two formulas to form a new database.

(1) Attribute abnormity: The importance of an attribute does not depend on the occurrence frequency of the attribute. Attribute importance is a specific value. The value is not more important when the value is greater, but if the value is quite different from the original value of the attribute, it indicates that there is a new rule contained in the new data, and this information is very important. Water quality monitoring network has detected the new situation; thus, it is more valuable for mining [20].

For numerical databases, we set the data elements in database D have m attributes that make up the set of attributes. The dependencies between these attributes are different, and the degree of importance is different as well. For the k-th attribute of an element, its importance is marked as E_k according to its importance [21]. This paper simplifies the calculation of the importance of attributes, attribute importance E_k can be calculated as follows:

$$E_k = \frac{\sum_{i=1}^{|D|} T_{ik}}{|D|} \tag{1}$$

Where T_{ik} is the k-th attribute in the i-th record. $T_{ik} \in \{0,1\}$; $|D|$ is the number of samples in database D.

(2) Difference of data set: the degree of difference $dif(D1, D2)$ between the original data set D_1 and the new data set D_2 is calculated via following formula:

$$dif(D_1, D_2) = \sqrt{\frac{\sum_{i=1}^{m}(E_{1i} - E_{2i})^2}{m}} \tag{2}$$

Where E_{1i} is the importance of the i-th attribute in D_1; E_{2i} is the importance of the i-th attribute in D_2; m is the total number of attributes in the entire data set.

Therefore, greater $dif(D1, D2)$ indicates greater difference between D_1 and the newly added data D_2, which means greater difference between the original data and the new data.

The main goal of this algorithm in the initial stage is to form a new database.

RULE UPDATING ALGORITHM PROCESS STEPS

The main steps of the rule updating algorithm are as follows.

(1) Obtain the degree of importance of each attribute $E1,E2,\ldots Em$ in D by (Formula 1);

(2) Obtain the degree of importance of each attribute $e_1,e_2,\ldots e_m$ in d by (Formula 1);

(3) Calculate the degree of difference $dis1 = dif(\mathrm{D}, \mathrm{d})$ between D and d by (Formula 2);

(4) Via $dis1$ from Formula (3), obtain the data randomly according to $dis1$;
When $0 < dis1 < 0.5$, choose $(1 - dis1) \times |\mathrm{D}|$ records from D to form D′;
When $0.5 < dis1 < 1$, choose $dis1 \times |\mathrm{D}|$records from D to form D′;

(5) Calculate the difference $dis2$ between D and D′,
When $dis2 < 0.2$, a new dataset Dd was formed with D and D′,
When $dis2 \geq 0.2$, return to the previous step and randomly choose new data again.

(6) Set evolutionary algebra T, with a group size of M;

(7) Initialize $t = 0$, select initial population G randomly from Dd;

(8) Calculate the fitness of individuals, and also calculate the concentration, crossover probability and mutation probability. Add rules that are greater than the minimum threshold to the association rules table.

(9) Through the selection, crossover and mutation operation, get the next generation of population;

(10) If $t < T$, then $t < \leftarrow t + 1$, return to the previous step; Otherwise, proceed the next step;

(11) Output the current mining result from the association rules table.

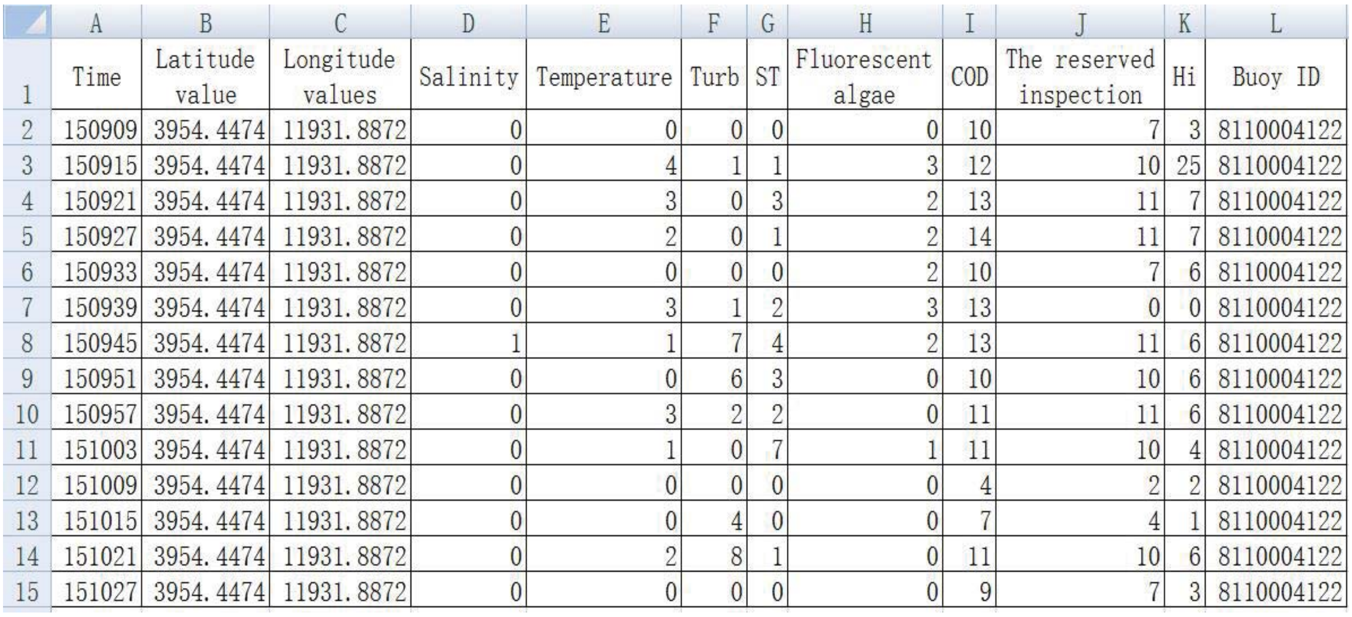
APPLICATION OF THE RULE UPDATING ALGORITHM IN THE WATER QUALITY MONITORING DATA MINING

In the actual rule updating algorithm application, the original database is generally large, and the incremental database is relatively small, if we do not use the rule updating method, each time the database changes after the mining, it will mine the mined original database once again, and it is a big waste of time. If we do adopt the rule updating method, and fully use the mining results of the original database, then when re-mining new data and part of the original data, the mining time will be much shorter, thus improve the mining efficiency. In the water quality monitoring database, under a relatively stable environment, the new data will not change greatly compared with the data in the original database. The values of each parameter are approximate and the results of incremental mining are also similar to the original results. However, when there is a big change in the marine ecological conditions, except for seasonal reasons, such as aviation fuel leaks and the sudden emission of pollutants and other emergencies, data with large differences would suddenly appear in the smooth monitoring data, thus will lead to big difference between the new data and the original data, in which case the algorithm needs to be re-standardized and re-mined, and it cannot reflect the advantages of the incremental mining algorithm. As the season changes, the performance of the photoelectric sensor for water quality monitoring may change in a stable working environment. The data in the database should be checked at regular intervals to ensure the accuracy of data mining.

The monitoring value of each field is divided into 1 ~ n according to different value intervals. Here, each field can be added 0 encoding, which represents that there is no relation between this attribute and other attributes. Randomly generate the rules under constraints. According to the coding of the specific data of the detected parameters, the parameters are mined to obtain the mapping table shown in Table 1.

CONCLUSION

When the original database obtains new data, the rules mining of the original database will have some failures, the new data brings new rules at the same time, and the overall



	A	B	C	D	E	F	G	H	I	J	K	L
1	Time	Latitude value	Longitude values	Salinity	Temperature	Turb	ST	Fluorescent algae	COD	The reserved inspection	Hi	Buoy ID
2	150909	3954.4474	11931.8872	0	0	0	0	0	10	7	3	8110004122
3	150915	3954.4474	11931.8872	0	4	1	1	3	12	10	25	8110004122
4	150921	3954.4474	11931.8872	0	3	0	3	2	13	11	7	8110004122
5	150927	3954.4474	11931.8872	0	2	0	1	2	14	11	7	8110004122
6	150933	3954.4474	11931.8872	0	0	0	0	2	10	7	6	8110004122
7	150939	3954.4474	11931.8872	0	3	1	2	3	13	0	0	8110004122
8	150945	3954.4474	11931.8872	1	1	7	4	2	13	11	6	8110004122
9	150951	3954.4474	11931.8872	0	0	6	3	0	10	10	6	8110004122
10	150957	3954.4474	11931.8872	0	3	2	2	0	11	11	6	8110004122
11	151003	3954.4474	11931.8872	0	1	0	7	1	11	10	4	8110004122
12	151009	3954.4474	11931.8872	0	0	0	0	0	4	2	2	8110004122
13	151015	3954.4474	11931.8872	0	0	4	0	0	7	4	1	8110004122
14	151021	3954.4474	11931.8872	0	2	8	1	0	11	10	6	8110004122
15	151027	3954.4474	11931.8872	0	0	0	0	0	9	7	3	8110004122

Fig. 1. Some monitoring data from the optical sensor network

Tab. 1. Field mapping table

ID	SI	TEMP	SAL	FA	COD	pH	TURB
150908	1	1	1	1	1	2	1
150909	2	5	1	3	3	2	2
150910	2	4	1	2	4	1	1
150911	1	3	1	2	5	1	1
150912	1	1	1	2	1	2	1
150913	2	4	1	3	4	1	2
150914	1	2	2	2	4	3	8
150915	2	2	1	2	2	1	3
150916	1	3	1	3	5	2	3
150917	5	1	1	2	2	1	2
150918	2	2	2	3	4	3	3
150919	1	1	2	2	2	3	2
150920	1	1	2	3	5	3	3
150921	2	1	1	2	3	2	2
150922	4	1	1	1	2	1	1
150923	1	4	1	1	2	3	2

remining of the new database is low in efficiency, so the rule updating mining algorithm is quite important. According to the characteristics of the water quality monitoring database studied in this paper, we focus on the numerical attribute database and propose a rule updating algorithm to solve the problem of rule maintenance caused by the dynamic changes of the database. According to the rule, this algorithm forms a new database from part of the original data and the new data, and searches the new database by random search and mining, which can quickly mine the effective rules without generating a large number of redundant rules. The experimental results show that the algorithm can quickly and effectively update the rules mining in the water quality monitoring database and can rapidly excavate potential information of the new data, which is of high practicability.

Tab. 2. Water quality monitoring data mining comparative result

Static database mining results	Rule updating mining results
0200010	0200010
0030003	0031013
0100110	0100110
0303000	0303000
0020300	0030300
3000010	3100010
4010000	4010000
0301000	0301000
0303000	0303000
0021000	0021001
1200000	1200001
0003003	0003013
1000110	1000110

ACKNOWLEDGMENT

Hebei Province Postdoctoral Research Project Funding No: B2018003012

REFERENCES

1. J. Wang, F. Zhang, and F. Liu, *Hybrid Forecasting Model-based Data Mining and Genetic Algorithm-adaptive Particle Swarm Optimisation: A Case Study of Wind Speed Time Series*, IET Renewable Power Generation, Vol. 10, No. 3, pp. 287–298, 2016.

2. C. K. Huynh, and W. C. Lee, *An Interference Avoidance Method using Two-Dimensional Genetic Algorithm for Multicarrier Communication Systems*, Journal of Communications and Networks, Vol. 15, No. 5, pp. 486–495, 2013.

3. H. Ghorbaninejad, and R. Heydarian, *New Design of Waveguide Directional Coupler using Genetic Algorithm*, IEEE Microwave and Wireless Components Letters, Vol. 26, No. 2, pp. 86–88, 2016.

4. M. M. Abouelsaad, M. A. Abouelatta, and A. R. Salama, *Genetic Algorithm-optimised Charge Simulation Method for Electric Field Modelling of Plate-Type Electrostatic Separators,* IET Science, Measurement & Technology, Vol. 7, No. 1, pp. 16–22, 2013.

5. B. Mohammadi-Ivatloo, A. Rabiee, and A. Soroudi, *Nonconvex Dynamic Economic Power Dispatch Problems Solution using Hybrid Immune-Genetic Algorithm*, IEEE Systems Journal, Vol. 7, No. 4, pp. 777–785, 2013.

6. W. H. Ip, D. Wang, and V. Cho, *Aircraft Ground Service Scheduling Problems and Their Genetic Algorithm with Hybrid Assignment and Sequence Encoding Scheme*, IEEE Systems Journal, Vol. 7, No. 4, 649–657, 2013.

7. S. H. Chung, and H. K. Chan, *A Two-Level Genetic Algorithm to Determine Production Frequencies for Economic Lot Scheduling Problem*, IEEE Transactions on Industrial Electronics, Vol. 59, No. 1, pp. 611–619, 2011.

8. W. Verly, L. R. Araujo, and D. R. R. Penido, *A Method for Sizing of Industrial Electrical Systems using Genetic Algorithm,* IEEE Latin America Transactions, Vol. 14, No. 2, pp. 681–686, 2016.

9. Y. Tominaga, Y. Okamoto, and S. Wakao, *Binary-based Topology Optimization of Magnetostatic Shielding by a Hybrid Evolutionary Algorithm Combining Genetic Algorithm and Extended Compact Genetic Algorithm*, IEEE Transactions on Magnetics, No. 49, No. 5, pp. 2093–2096, 2013.

10. T. Lu, and J. Zhu, *Genetic Algorithm for Energy-Efficient QoS Multicast Routing,* IEEE Communications Letters, Vol. 17, No. 1, 31-34, 2012.

11. L. Shi, Y. K. Deng, and H. F. Sun, *An Improved Real-Coded Genetic Algorithm for the Beam Forming of Spaceborne SAR*, IEEE Transactions on Antennas and Propagation, Vol. 60, No. 6, pp. 3034–3040, 2012.

12. K. Boudjelaba, F. Ros, and D. Chikouche, *Adaptive Genetic Algorithm-based Approach to Improve the Synthesis of Two-Dimensional Finite Impulse Response Filters*, IET Signal Processing, Vol. 8, No. 5, pp. 429–446, 2014.

13. T. A. Mota, J. F. Leal, and A. C. Lima, *Neural Equalizer Performance Evaluation using Genetic Algorithm*, IEEE Latin America Transactions, Vol. 13, No. 10, pp. 3439–3446, 2015.

14. K. Thirugnanam, M. Singh, and P. Kumar, *Mathematical Modeling of Li-Ion Battery using Genetic Algorithm Approach for V2G Applications*, IEEE Transactions on Energy Conversion, Vol. 29, No. 2, pp. 332–343, 2014.

15. X. K. Wei, W. Shao, and C. Zhang, *Improved Self-Adaptive Genetic Algorithm with Quantum Scheme for Electromagnetic Optimisation*, IET Microwaves, Antennas and Propagation, Vol. 8, No. 12, pp. 965–972, 2014.

16. S. C. Sen, S. Kasim, M. F. Md Fudzee, R. Abdullah, and R. Atan, *Random Walk from Different Perspective*, Acta Electronica Malaysia, Vol. 1, No. 2, pp. 26–27, 2017.

17. B. Q. Li, and Z. Li (). *Design of Automatic Monitoring System for Transfusion*, Acta Electronica Malaysia, Vol. 2, No. 1, pp. 07–10, 2018.

18. X. N. Gu, Z. G. He, X. Y. Sun, J. Liu, and B. S. Wang, *Algebraic dynamic multilevel (ADM) method for compositional multi-phase flow simulation*, Acta Mechanica Malaysia, Vol. 1, No. 1, pp. 01–03, 2017.

19. X. Luo, *Research on Anti-Overturning Performance of Multi-Span Curved Girder Bridge with Small Radius*, Acta Mechanica Malaysia, Vol. 1, No. 1, pp. 11–15.

20. S. K. Md Sa'at, and N. Qamaruzaman, *Phytoremediation potential of palm oil mill effluent by constructed wetland treatment*, Engineering Heritage Journal, Vol. 1, No. 1, pp. 49–54, 2017.

21. S. R. Hassan, N. Qamaruz Zaman, I. Dahlan, *Influence of Seed Loads on Start Up of Modified Anaerobic Hybrid Baffled (MAHB) Reactor Treating Recycled Paper Wastewater*, Engineering Heritage Journal, Vol. 1, No. 2, pp. 05–09, 2017.

CONTACT WITH THE AUTHORS

Junhai Zhang
e-mail: zhangresc@163.com

Postdoctoral research station of Geography
Hebei Normal University
Shijiazhuang 050024
China