Minimum Monotonous Constraint Closure Hadoop Parallel Association Rules Under Big Data Environment

Jin Ou[1], Jin Yiqiao[1], He Jianbiao[1], Li Xi[1]

1. School of Information Science and Engineering, Central South University, Hunan Changsha, 410083, China

Key Word: Big Data; Closure Operator; Minimum Monotonous Constraint; Hadoop Framework

Abstract: Aiming at the lager rule redundancy problems in traditional association rules, this article proposes minimum monotonous constraint closure Hadoop parallel association rules. First, basing on closure operator constraint rule equivalence relation set, this article gives satisfying minimum monotonous constraint rule set which can effectively divide the constraint rule set into disjoint equivalence rule class to reduce the rate of redundancy rule. Second, aiming at the big data problems, this article adopts Mapreduce parallel computation model under Hadoop framework to realize the parallelization computation of minimum monotonous constraint association rules which effectively promote the expansibility of algorithm to big data treatment. At last, through experimental comparison on standard test set, this article shows the effectiveness of the proposed algorithm.

Correlation Studies

In order to reduce the storage and computation time of big data, the technology of reducing data mining algorithm redundancy basing on constraint limit has obtained extensive research and application. In the initial stage of studying this technology, the research direction is to use original constraint for the data mining algorithm. The most typical example is to reduce redundancy of discovered frequent item set through finding the constraint with lowest frequency in the transaction database, and then implement constraint to the frequent item set of association rules through minimum confidence degrees to effectively reduce association rules algorithm redundancy and increase algorithm efficiency.

For model $T_m = (O, A, R)$, when using traditional association rules for computation, if the T_m value is large, then the algorithm has a low operation efficiency. In addition, only implementing constraint with confidence degrees and support degrees, the users cannot locate interest subset quickly and the algorithm efficiency is low. So scholars propose a more complicated constraint limit to simulate the real demands of users to further reduce cost of algorithm. For example, literature [4] designed monotonous and anti-monotonous characteristic constraint which are C_{m} and C_{am} respectively. Basing on the above research achievements, this article also proposes to adopt association rules equivalence relation set of closure constraint rules from algorithm constraint point, comprehensively considers maximum confidence degrees and support degrees threshold value as well as minimum monotonous constraint which implement disjoint division of equivalence rules class to constraint rules set $\mathrm{ARS}_{\supseteq L_0, \supseteq R_0}(s_0, s_1, c_0, c_1)$.

Problem Description of Association Rules

For model $T_m = (O, A, R)$, O is non-empty object constraint. A is object characteristic attribute, R is binary relation set $O \times A$. The base numbers of O and A can be indicated as $n = |O|$ and $m = |A|$. $X \subseteq A$ is item set whose support set is $\mathrm{supp}(X)$. Let the upper and lower limit of support threshold values as s_0 and s_1 respectively. Among them, $0 < 1/n \le s_0 \le s_1 \le 1$, $n = |O|$. If A is non-empty

frequency item set and satisfies the condition $s_0 \le \sup p(A) \le s_1$, then non-empty proper subset L' can be extracted from any frequency item set S' and $\varnothing \ne L' \subset S'$, $R' \equiv S'/L'$. Establish rule relation $L' \to R'$ basing on subset L', the computation mode of confidence degree and support degree of R' is:

$$\begin{cases} \mathrm{supp}(r) \equiv \mathrm{supp}(L') \\ \mathrm{conf}(r) \equiv \mathrm{supp}(S')/\mathrm{supp}(L') \end{cases} \tag{1}$$

The upper limit and lower limit of confidence threshold values are c_0 and c_1 respectively and satisfy the condition $0 < c_0 \le c_1 \le 1$, the under the conditions of $c_0 \le \mathrm{conf}(r)$ and $s_0 \le \mathrm{supp}(r)$, the association rules set is

$$\mathrm{ARS}(s_0, c_0) \equiv \left\{ \begin{array}{l} r : L' \to R' \left| \varnothing \ne L', R' \subseteq A \right. \\ L' \cap R' = \varnothing,\ S' \equiv L' + R' \\ s_0 \le \mathrm{supp}(r), c_0 \le \mathrm{conf}(r) \end{array} \right\} \tag{2}$$

In the formula, $\mathrm{ARS}(s_0, c_0)$ is association rule. The study of questions in association rules are most concentrated on confidence, subset and multiple constraint support aspects. The additional constraint of association rules has dual characters which are L_0 and $R_0 \subseteq A$. The goal of algorithm is to have discoveries to all the existing rules $r : L' \to R'$, then its support degree and confidence degree satisfy conditions $s_0 \le \mathrm{supp}(r) \le s_1$ and $c_0 \le \mathrm{conf}(r) \le c_1$ as well as item set constraint $L' \supseteq L_0$ and $R' \supseteq R_0$. The above minimum monotonous constraint can be specifically expressed as:

$$\mathrm{ARS}_{\supseteq L_0, \supseteq R_0}(s_0, s_1, c_0, c_1) \equiv \{ r : L' \to R' \in \mathrm{ARS}(s_0, s_1, c_0, c_1) \left| L' \supseteq L_0, R' \supseteq R_0 \right. \} \tag{3}$$

Among it:

$$\mathrm{ARS}(s_0, s_1, c_0, c_1) \equiv \left\{ \begin{array}{l} r : L' \to R' \in \mathrm{ARS}(s_0, c_0) \\ \left| \mathrm{supp}(r) \le s_1, \mathrm{conf}(r) \le c_1 \right. \end{array} \right\} \tag{4}$$

If $s_1 = c_1 = 1$, $L_0 = R_0 = \varnothing$, then the traditional association rules problems $\mathrm{ARS}(s_0, c_0)$ can be obtained.

Minimum Monotonous Constraint Closure Association Rules

Coarse Division

In order to further reduce candidate solution redundancy, set disjoint equivalence rule partition and use closure operator to design disjoint equivalence relation set of frequent item sets $\mathrm{FS}(s_0, s_1)$ and $\mathrm{ARS}(s_0, s_1, c_0, c_1)$.

Definition $\mathrm{FS}(s_0, s_1)$ and $\mathrm{ARS}(s_0, s_1, c_0, c_1)$ have equivalence relation sets.

(1) $\forall A', B \in \mathrm{FS}(s_0, s_1)$, $B \Leftrightarrow h(A) = h(B)$, $A' \subseteq A$;

(2) $\forall r_k : L_k \to R_k \in \mathrm{ARS}(s_0, s_1, c_0, c_1)$, $r_2 \Leftrightarrow [h(L_1) = h(L_2), h(L_1 + R_1) = h(L_2 + R_2)]$, $k = 1, 2$ $r_1 \subseteq r$.

Let $\mathrm{FCS}(s_0, s_1) = \mathrm{FS}(s_0, s_1) \cap \mathrm{CS}$ as closure operator, then for $\forall L \in \mathrm{FCS}(s_0, s_1)$, there is equivalence relation $[L]_A \equiv \{ \varnothing \ne L' \subseteq L, h(L') = L \}$. The meaning is the equivalence frequent item set class of same closure L. $\forall L, S \in \mathrm{FCS}(s_0, s_1)$, $\varnothing \ne L \subseteq S$, $\mathrm{supp}(S)/\mathrm{supp}(L) \in [c_0, c_1]$, for all the rule equivalence class $r : L' \to R'$, then $h(L') = L$ and $h(L' + R') = S$ can be expressed as:

$$\begin{array}{l} \mathrm{AR}(L, S) \equiv \{ r : L' \to R' \in \mathrm{ARS}(s_0, s_1, c_0, c_1) \\ \left| L' \in [L]_A, S' \equiv L' + R' \in [S]_A \right. \} \end{array} \tag{5}$$

Deduction: Coarse Division: the coarse division process of association rule constraint rule can be expressed in the following formula:

$$\mathrm{ARS}_{\supseteq L_0, \supseteq R_0}(s_0, s_1, c_0, c_1) = \sum_{(L,S) \in \mathrm{NFCS}(s_0, s_1, c_0, c_1)} \mathrm{AR}_{\supseteq L_0, \supseteq R_0}(L, S) \tag{6}$$

$$\mathrm{AR}_{\supseteq L_0, \supseteq R_0}(L, S) \equiv \{ r : L' \to R' \in \mathrm{AR}(L, S) \left| L' \supseteq L_0, R' \supseteq R_0^{(t)} \right. \}.$$

Smooth Partition Minimum Monotonous Constraint

Theorem the constraint rule of smooth partition is if the limitation of rule satisfies condition H_1, then there is:

$$\mathrm{ARS}_{\supseteq L_0,\supseteq R_0}(s_0,s_1,c_0,c_1)=\sum_{(L,S)\in \mathrm{NFCS}_{\supseteq L_0,\supseteq R_0}(s_0,s_1,c_0,c_1)} \mathrm{AR}^{+}_{\supseteq L_0,\supseteq R_0}(L,S) \tag{7}$$

The definition of the form of $\mathrm{NFCS}(s_0,s_1,c_0,c_1)$ is:

$$\begin{aligned}\mathrm{NFCS}(s_0,s_1,c_0,c_1)\equiv\{(L,S)\in \mathrm{CS}^2 \,|\, S\in \mathrm{FCS}(s_0,s_1),\\ \varnothing\neq L\subseteq S, \mathrm{supp}(S)/\mathrm{supp}(L)\in[c_0,c_1]\end{aligned} \tag{8}$$

Basing on the above rules and limitations, the candidate class $(L,S)\supseteq \mathrm{NFCS}_{\supseteq L_0,\supseteq R_0}(s_0,s_1,c_0,c_1)$ can be obtained from $\mathrm{NFCS}(s_0,s_1,c_0,c_1)$. But there is still a large amount of this kind of candidate class which means there are many redundant candidates. The purpose of the algorithm proposed by this article is to find frequent close item set $\mathrm{FCS}_{C_0\subseteq C_1}(s_0',s_1')$. In $\mathrm{FCS}_{C_0\subseteq C_1}(s_0',s_1')=\mathrm{MFCS\_FL}(\mathrm{LCG},S_0^*,A,s_0^*,s_1^*)$,

$\mathrm{LCG}\equiv\{(S,\mathrm{supp}(S),G(S))\,|\,(S,\mathrm{supp}(S))\in \mathrm{LC}\}$. The pseudo-code 1 of algorithm is as follows:

Pseudo-code 1: Smooth Partition Constraint Rule

$\mathrm{FCS}_{C_0\subseteq C_1}(s_0',s_1')=\mathrm{MFCS\_FL}(\mathrm{LCG},C_0,C_1,s_0',s_1')$:

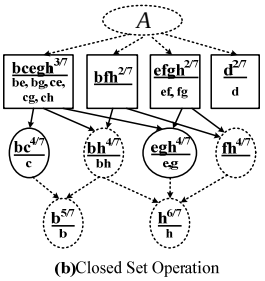
(1) $\mathrm{FCS}_{C_0\subseteq C_1}(s_0',s_1')$: $=\varnothing$;
(2) if $(s_0'>s_1')\,||\,(C_0\subseteq C_1)\,||\,(\mathrm{supp}(C_0)<s_0')\,||\,(s_1'<\mathrm{supp}(C_1))$ then
(3) return $\varnothing$;
(4) endif.
(5) for each $((L,\mathrm{supp}(L),G(L))\in \mathrm{LCG}^s)$ do
(6) if $(s_0'\le \mathrm{supp}(L)\le s_1' \,\&\, L\supseteq C_0)$ then
(7) if $(\exists L_i\in G(L)\,\&\, L_i\subseteq C_1)$ then
(8) $L_{C_1}=L\cap C_1$; $G_{C_1}(L)=\{L_i\in G(L)\,|\,L_i\subseteq C_1\}$;
(9) $\mathrm{FCS}_{C_0\subseteq C_1}(s_0',s_1'):=\mathrm{FCS}_{C_0\subseteq C_1}(s_0',s_1')+(L_{C_1},\mathrm{supp}(L_{C_1}))$;
(10). endif.
(11) endif.
(12) endfor.
(13) return $\mathrm{FCS}_{C_0\subseteq C_1}(s_0',s_1')$;

Specific Examples

The data set T_D of specific examples and corresponding closed item set are shown in Figure 1.

Trans	Items
1	bcegh
2	bcegh
3	bcd
4	defgh
5	bcefgh
6	bfh
7	fh

(a)Item Set List



(b)Closed Set Operation

Figure 1 Example Data Set

The maximum support degree chooses a threshold value $s_1=0.28$, the minimum support chooses a threshold value $s_0=0.28$. The upper limit and lower limit of confidence degree threshold values are $c_1=0.9$ and $c_0=0.4$ respectively.

(1)Constraint 1: $L_0=\mathrm{c}$, $R_0=\mathrm{f}$. First use PMM algorithm to generate $|\mathrm{ARS}(s_0,s_1,c_0,c_1)|=134$ association rules and then inspect them through limitations L_0 and R_0. For the 15 classes of rules of the above $\mathrm{NFCS}(s_0,s_1,c_0,c_1)$, there are $\mathrm{AR}(L,S)=\varnothing$, and

$\mathrm{ARS}_{\supseteq L_0,\supseteq R_0}(s_0,s_1,c_0,c_1)=\varnothing$.

(2) Constraint 2: $L_0=\mathrm{h}$, $R_0=\mathrm{b}$. Use the above PMM algorithm to generate 134 rules. Under the condition of obtaining 4 groups of class rules (L,S) of $\mathrm{NFCS}(s_0,s_1,c_0,c_1)$ through limitation inspection, there are $\left|\mathrm{ARS}_{\supseteq L_0,\supseteq R_0}(s_0,s_1,c_0,c_1)\right|=19$ rules with rule classifications of (egh,bcegh), (h,bcegh), (fh,bfh) and (h,bh) respectively. For the remaining $\left|\mathrm{NFCS}(s_0,s_1,c_0,c_1)\right|-4=11$ classifications, the algorithm will generate $\left|\mathrm{ARS}(s_0,s_1,c_0,c_1)\setminus\mathrm{AR}_{\supseteq L_0,\supseteq R_0}(s_0,s_1,c_0,c_1)\right|=115$ candidate redundant rules, then $\mathrm{AR}_{\supseteq L_0,\supseteq R_0}(L,S)=\varnothing$. For rule class $(\mathrm{bc,bcegh})\in\mathrm{NFCS}(s_0,s_1,c_0,c_1)$, then 21 rule candidates can be further obtained basing on PMM algorithm.

(3)Constraint 3: $L_0=\mathrm{f}$, $R_0=\mathrm{h}$. There are only 4 rules $(L^1=\mathrm{fh},S^1=\mathrm{efgh})$, $(L^2=\mathrm{fh},S^2=\mathrm{bfh})$ satisfy condition $\mathrm{AR}_{\supseteq L_0,\supseteq R_0}(L^i,S^i)=\varnothing$, $i=1,2$, then for $(L^1=\mathrm{fh},S^1=\mathrm{efgh})$, then the amount of rule candidates generated by $\mathrm{AR}(L^1,S^1)$ is 9.

Experiment and Analysis

In order to evaluate the performance of proposed algorithm, use Hadoop 1.0.4 and Ubuntu 12.10 to establish calculation clusters of 10 sets of machines (one of them is management mainframe). Every machine is allocated with 2.0 GHz double kernal CPU, 4G internal storage and 320G hard disk. Compare algorithms to select parallel Apriori algorithm (Parallel Apriori Algorithm，PAA)and AprioriPMR algorithm. The simulation software selects Matlab2012b. In order to make experiment to compare performances of CMSC-HPAR algorithm, Apriori algorithm and AprioriPMR algorithm, select the algorithm with a computational node change between 1-9 to implement time change for comparison.

Set three data sets D_1、D_2、D_3. There are 5 data documents in D_1 with a total 0.75GB. The average affair has 20 item sets. There are 5 data documents in D_2 with a total 1.0 GB. The average affair has 40 item sets. There are 5 data documents in $D_3$with a total 1.5 GB. The average affair has 20 item sets. These three set aims at different amounts of data documents, different amount of average affair item sets, and different total size of general documents. The purpose of establishment through the 3 aspects is to simulate data environment under different conditions as possible as one can to verify performance of algorithm more adequately. This article's support degree threshold value is a smaller value (0.05%). The experiment comparison results on D_1 data set is shown in Figure 2.

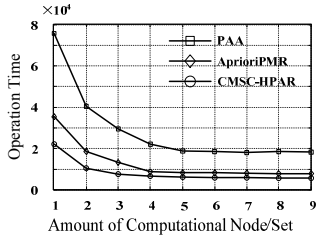


Figure 2 Experiment Result of D_1 Data Set

It can be seen from the figure that the operation time of CMSC-HPAR algorithm is smallest. In the meantime, it can also be seen that when there are smaller amount of data size and affair item sets and after computational node achieves $n=3$, the promotion of the parallel node number to computation capability achieves saturation gradually. Any more computational node number has limited capability to reduce operation time.

Conclusion

Aiming at the redundant problems of traditional association rule algorithm, this article proposes minimum monotonous constraint closure association rule algorithm. In order to increase the computational efficiency of algorithm, basing on Mapreduce parallel computation mode under Hadoop framework, this realizes parallelization computation of minimum monotonous constraint closure association rule. It is shown through experiment comparison results that the computational efficiency of proposed algorithm is better than compared algorithm which shows it expansibility in treating big data and has higher practical application value. The following study includes: one, further optimize algorithm structure; two, consider to combine it with other parallel computation mode to discuss advantages and disadvantages of different algorithms which can provide guidance to practical application.

Reference

[1] Zhihan Lv, Tengfei Yin, Yong Han, Yong Chen, and Ge Chen. WebVR——web virtual reality engine based on P2P network. Journal of Networks. 6, no. 7 (2011): 990-998.

[2] Jiachen Yang, Bobo Chen, Jianxiong Zhou, Zhihan Lv. A portable biomedical device for respiratory monitoring with a stable power source. Sensors. 2015.

[3] Zhao, Dongfang, et al. "FusionFS: Toward supporting data-intensive scientific applications on extreme-scale high-performance computing systems." Big Data (Big Data), 2014 IEEE International Conference on. IEEE, 2014.

[4] Shuping Dang, Jiahong Ju, Matthews, D., Xue Feng, Chao Zuo. Efficient solar power heating system based on lenticular condensation. Information Science, Electronics and Electrical Engineering (ISEEE), 2014 International Conference on . 26-28 April 2014