ATLANTIS PRESS
Advances in Intelligent Systems Research (AISR), volume 150
2017 International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2017)

Flow Graph Network Based Non-redundant Correlative Educational Rules Discovered

Bo Liu[1,a], Changqin Huang[1,b], Xiuyu Lin[1,c]

[1]School of Education Information Technology, South China Normal University, Guangzhou, China

[a]gavinliu2003@126.com, [b]cqhuang@zju.edu.cn, [c]free20@126.com

Keywords: correlative rule; non-redundant rule; flow graph network

Abstract: A correlative rule expresses a relationship between two correlative events happening one after another. These rules are potentially useful for analyzing correlative data, ranging from purchase histories, web logs and program execution traces. In this work, we investigate and propose a syntactic characterization of a non-redundant set of correlative rules built upon past work on compact set of representative patterns. When using the set of mined rules as a composite filter, replacing a full set of rules with a non-redundant subset of the rules does not impact the accuracy of the filter. Lastly, we propose an algorithm to mine this compressed set of non-redundant rules. A performance study shows that the proposed algorithm significantly improves both the run-time and compactness of mined rules over mining a full set of sequential rules.

1 Introduction

In recent years association rules mining has been studied more and more widely in the field of data mining [1, 2]. Z.Pawlak proposes a new approach to knowledge representation and data mining based on flow analysis in a new kind of flow network [3]. Bo L. presented a data-processing model [4, 5, 6] based on non-linear correlation discovery (NLCD), which Using non-linear technique to analysis and quantify the correlation between the attribute groups; Adopting the data presentation of bool serial, this method can map the original data into bool serials and get the easily processing object data; Applying the holistic correlation analysis into the non-linear analysis and quantifying of the correlation; Using non-linear technique to settle those problems above, and the experiment evaluation showing our model NLCD is efficient. David Lo, SiauChe-ng Khoo, Limsoon Wong propose and characterize a non-redundant set of sequential rules [7]. We have made deep researching of the roles in the asynchronous collaborative learning network, finding that there were many kinds of correlative interaction between them[8].

While the correlation between pair-wise attribute groups can be mined by NLCD, there were most directions of these correlations that are very difficult to discover by the normal statistic technique through can not be ignored. At the same time, we should be aware how these correlations are working, i.e. knowing how and the degree of these correlations. But these problems cannot be solved through model of NLCD.

To tame the explosive growth of rules, we propose mining a non-redundant set of correlative rules. Central to our method is the notion of rule inference. This notion is used to define and remove redundancy among rules. When using the set of mined rules as a composite filter, replacing a full set of rules with the non-redundant subset of rules does not impact the accuracy of the filter.

We propose an algorithm to mine this compressed set of non-redundant correlative rules. Our performance study shows much benefit in mining non-redundant correlative rules over a full set of rules. The study shows that the runtime and number of rules mined can be reduced by up to 31 times and 21 times compared with NLCD algorithm [6], respectively.

The contributions of our work are as follows:

(1)We propose a concept of non-redundant correlative rules based on flow graph network rule inferring.

(2) We investigate different sets of patterns and their various compositions to form different sets of rules. We study the quality of these rule sets with respect to completeness and tightness.

(3) We characterize a tight and complete set of non-redundant correlative rules based on compositions of patterns.

(4) We propose and characterize compression of the non-redundant set of correlative rules.

(5) We develop an algorithm to mine the compressed set of non-redundant correlative rules and show that it performs much faster than mining a full set of correlative rules.

The rest of this paper is organized as follows. Section 2 provides a detailed describing of terminologies and definitions used. Section 3 introduces the correlative rules referring and its reduction algorithm(NODARF, NOn-reDundant correlAtive Rules reFerring) . Section 4 shows a detailed experimental evaluation of NODARF on datasets from different domains and compares it against existing algorithms. Finally, Section 5 provides concluding remarks and the works in the future.

2 Preliminaries

2.1Non-redundant correlative rule

Here we give the criterion notation on the association rules mining: let $I=\{i_1,i_2,..,i_m\}$ be the itemset, and the D be a database of transactions, a transaction T containing itemset I if and only if $I \subset T$. The support of an itemset I, denoted $\sigma(I)$, is the number of transactions in which it occurs an as subset. Then association rule is a condition of the form $A \rightarrow B(p, q)$, where $A \subset I$， $B \subset I$ and $A \bigcap B=\phi$, if this rule is true in D when it satisfies two conditions, such as support and confidence.

A rule is said to be redundant in a set of rules R iff it can be inferred by another rule in R.

Consider the following two rules: $r_1=\{A\} \rightarrow \{B; C; D\}$ and $r_2 = \{A\} \rightarrow \{B\}$ having the same support and confidence. r_2 is redundant since it can be inferred by r_1.

2.2 Flow graph network

A flow graph network is a directed, acyclic, finite graph G = (V, E, f), where V is a set of nodes, $E \subseteq V \times V$ is a set of directed edges, $f: E \rightarrow R^+$ is a flow function and R^+ is the set of non-negative real.

Input set of a node $x \in V$ is the set $I(x) = \{y \in V: (y, x) \in E\}$; output set of a node $x \in V$ is defined by $O(x) = \{y \in V: (x, y) \in E\}$. We will also need the concept of input and output of a graph G, defined, respectively, as follows: $I(G) = \{x \in V : I(x) = \varnothing\}$, $O(G) = \{x \in V : O(x) = \varnothing\}$. Inputs and outputs of G are external nodes of G; other nodes are internal nodes of G. If $(x, y) \in E$, then f(x, y) is a through flow from x to y.

With every node x of a flow graph G we associate its inflow

$$f_+(x) = \sum_{y \in I(x)} f(y,x) \qquad (1)$$

and outflow

$$f_-(x) = \sum_{y \in O(x)} f(x,y) \qquad (2)$$

Similarly, an inflow and an outflow for the whole flow graph are defined by

$$f_+(G) = \sum_{x \in I(G)} f_-(x) \qquad (3)$$

$$f_-(G) = \sum_{x \in O(G)} f_+(x) \qquad (4)$$

We assume that for any internal node x we have $f_+(x) = f_-(x) = f(x)$, where $f(x)$ is a through-flow of node x.

Then, obviously, $f_+(G) = f_-(G) = f(G)$, where $f(G)$ is a through-flow of graph G.

With every edge (x, y) of a flow graph G the certainty and the coverage factors are associated. The certainty (Cer(x,y)) and the coverage (Cov(x,y)) of (x, y) are defined by

$$\text{Cer(x,y)} = \frac{\sigma(x,y)}{\sigma(x)} \qquad (5)$$

$$\text{Cov(x,y)} = \frac{\sigma(x,y)}{\sigma(y)} \qquad (6)$$

Here $\sigma(w)$ is a normalized factor of w, for example, $\sigma(x,y)=\frac{f(x,y)}{f(G)}$ being the normalized flow of edge (x, y); $\sigma(x)=\sigma_{+}(x) = \frac{f_{+}(x)}{f(G)}=\frac{f_{-}(x)}{f(G)}=\sigma_{-}(x)$ being the normalized through-flow of x.

A (directed) path from x to y, $x \neq y$ in G is a sequence of nodes $x_1, \cdots\cdots, x_n$ such that $x_1 = x$, $x_n = y$ and $(x_i, x_{i+1}) \in E$ for every i, $1 \le i \le n-1$. A path from x to y is denoted by [x ⋯ ⋯ y].The certainty of the path $[x_1 \cdots \cdots x_n]$ is defined by

$$\text{Cer}[x_1 \cdots \cdots x_n] = \prod_{i=1}^{n-1} cer(x_i, x_{i+1}) \qquad (7)$$

the coverage of the path $[x_1 \cdots \cdots x_n]$ is defined by

$$\text{Cov}[x_1 \cdots \cdots x_n] = \prod_{i=1}^{n-1} \text{cov}(x_i, x_{i+1}) \qquad (8)$$

the strength of the path $[x_1 \cdots \cdots x_n]$ is defined by

$$\begin{aligned}\text{Str}[x_1 \cdots \cdots x_n] &= \sigma(x_1) cer[x_1 \cdots x_n] \\ &= \sigma(x_n)\,\text{cov}[x_1, x_n] \end{aligned} \qquad (9)$$

The set of all paths from x to y ($x \neq y$) in G, denoted by <x, y >, will be called a connection from x to y in G. In other words, connection < x, y > is a sub-graph of G determined by nodes x and y. The certainty of the connection < x, y > is

$$\text{Cer<x, y>} = \sum_{[x \cdots y] \in <x,y>} cer[x \cdots y] \qquad (10)$$

the coverage of the connection < x, y > is defined by

$$\text{Cov<x, y>} = \sum_{[x \cdots y] \in <x,y>} \text{cov}[x \cdots y] \qquad (11)$$

the strength of the connection < x, y > is defined by

$$\begin{aligned}\text{Str<x, y>} &= \sum_{[x \cdots y] \in <x,y>} Str[x \cdots y] \\ &= \sigma(x) cer < x, y > = \sigma(y)\,\text{cov} < x, y > \end{aligned} \qquad (12)$$

If we substitute simultaneously any sub-graph < x, y > of a given flow graph G, where x and y are input and output nodes of G respectively, by a single branch (x, y) such that σ(x, y) = σ < x,y >, then in the resulting graph G', called the reduction of G, we have cer(x, y) = cer < x, y >, cov(x, y) =cov < x, y > and σ(G) = σ(G').

3 Rules referring

3.1 Non-redundant rules referring

Let us assume that the set of nodes of a flow graph network is interpreted as a set of objects considered. With every edge (x, y) we associate a correlative rule x → y, read if x then y; x will be referred to as condition, whereas y – decision of the rule. Such a rule is characterized by three parameters, σ (x, y), cer(x, y) and cov(x, y).

Let us observe that the inverted flow graph gives reasons for decisions. Every path $[x_1 \cdots x_n]$ determines a sequence of decision rules $x_1 \rightarrow x_2$, $x_2 \rightarrow x_3$, $\cdots$, $x_{n-1} \rightarrow x_n$. From previous considerations it follows that this sequence of decision rules can be interpreted as a single decision rule $x_1x_2 \ldots x_{n-1} \rightarrow x_n$, in short $x_* \rightarrow x_n$, where $x_* = x_1x_2 \ldots x_{n-1}$, characterized by

$$\mathrm{Cer}(x_*,x_n) = \frac{\sigma(x_*,x_n)}{\sigma(x_*)} \qquad (13)$$

$$\mathrm{Cov}(x_*,x_n) = \frac{\sigma(x_*,x_n)}{\sigma(x_n)} \qquad (14)$$

$$\mathrm{Str}(x_*,x_n) = \mathrm{Str}[x_1 \cdots x_n],\ \mathrm{Str}(x_*) = \mathrm{Str}[x_1 \cdots x_{n-1}] \qquad (15)$$

The set of all correlative rules $x_{i1}x_{i2} \cdots x_{in-1} \rightarrow x_{in}$ associated with all paths $[x_{i1} \ldots x_{in}]$ such that x_{i1} and x_{in} are input and output of the graph respectively will be called a non-redundant correlative rule induced by the flow graph.

3.2 Mining algorithm

To generate the non-redundant correlative rules set, one must first mine the basic rules from data, and then apply the reduction technique to them for the non-redundant correlative rules set. Our proposed algorithm (NODARF mining algorithm) is shown in Fig 1.

Procedure Mine Non-redundant Rules Algorithm
Input:CorDB: Correlative Database;
min_sup,min_conf: Minimum support and confidence thresholds
Outputs: *Rules*: Compressed set of non-redundant rules

Let *BS*={}, *CS*={},*Rules*={};
Concurrently mine BS and CS with support≥*min_sup*;
For each rule r *BS*
  Let *Related*={r' ∈CS,where $r \rightarrow r'$ and $\min\_conf \le \sup(r')/\sup(r) \le 1$};
Let *Rules*'={};
For each rule *r*' ∈*Related*
  Let *post=px*, where *sx++px=r*', *sx*→*r*, and *sx* is as short as possible;
    Let *rn*=r→post ,*nsup*=sup(r') and *nconf*= sup(r')/sup(r);
      if ∃ r(s,c) ∈*Rules*', such as *r*.post=*rn*.post;
        if(*nsup* > *s)*Replace *r(s,c)* in *Rules*' with *r(nsup, nconf)*;
      else if ∃ *r(s,c)* ∈Rules', such as *r*.post →*rn*.post
        if(*nsup* > *s*)Remove *r(s,c)* from *Rules*';
          Add *rn(nsup,nconf)* to *Rules*';
      else Add rn(*nsup,nconf*) to *Rules*';
Rules=Rules Rules';
Output *Rules*

Figure 1. NODARF algorithm.

Table 1 Experiment data

data	Times (days)	Attribute number	TID number
D070101	180	8	14532
D060312	360	5	32231

Table 2 data

TID	ATTRIBUTES
100	A1,A2,A3,A4,A5,A6,A7,A8
200	A1,A2,A3,A4,A5,A6,A7,A8
300	A2,A4,A5,A6,A8
400	A1,A3,A4,A5,A7,A8
500	A3,A4,A5,A6,A7,A8

4 Experimental results

In this paper our experiment environment is in the personal computer of Intel(R) Pentium(R) 4, 1843 MHz-CPU、DDR512MB-memory、Win-XP system and all programs are compiled by C++. We make use of data sets from SCNUoL system showing in the table 1, 2, 3. Here we run our algorithm NODARF and talk about the experiment results (showing in the following table 4,5). To test the performance of it, the other experiment based on NLCD is run compared with NODARF.

Firstly we pre-processed data above, for example seeing the table 2. Where TID meaning the mark of each log, ATTRIBUTES meaning the attributes of the logs atrributes(including A1-A8).

After pretreatment of the data, we can get the correspond -ing symbol sequences(seeing table 3)based on the eight atrributes, the process done later mostly based on this table.

4.1 First experiment

At the first experiment we will consider the relation among the attributes values of A2 and A8. As showing in table 4, four remarkable rules were selected here from about 138 rules discovered by

Table 3 Symbol table

Atrributes	Symbol
A1	Id
A2	A: assignment, Q: quiz, R: read, D: discussing
A3	Courseid
A4	M: morning, N: afternoon, V: evening
A5	M: morning, N: afternoon, V: evening
A6	Num
A7	String
A8	F: fall, P: pass, G: good, E: excellent

NODARF, from which three rules described that most students who had finished reading and discussing and at least passing the course would refer assignments including 1, 2 and 3, however those who fell in the course had only referred assignment 1 from the last rule.

Table 4 Experiment report

R∧D∧E∧V==>e1 [sup:81% conf:83%]
C∧R∧P==>e2 [sup:77% conf:86%]
E∧R∧D∧P==>e3 [sup:69% conf:71%]
R∧D∧F==>e1 [sup:59% conf:89%]

e1: people referring assignment 1;
e2: people referring assignment 2;
e3: people referring assignment 2;

4.2 Second experiment

At the second experiment we will consider the connection among three different kinds of assignments e1, e2 and e3. As showing in table 5, assignment e1 and e3 has the much stronger correlation than other two relations: e1 and e2, e2 and e3. So the instructor can design correlative assignments to motivate students in time to learn the course.

Table 5 Experiment report

e1==>e2 [sup:33% conf:75%]
e2==>e1 [sup:38% conf:83%]
e2==>e3 [sup:60% conf:79%]
e3==>e2 [sup:61% conf:81%]
e1==>e3 [sup:76% conf:90%]
e3==>e1 [sup:76% conf:89%]

From two experiments above, the NODARF algorithm is compared with NLCD based on runtime used and rules found in them. The study shows large improvements in both runtime and compactness of minded rules (seeing fig 2, 3, 4, 5 in Appendix A), runtime being improved up to 31 times and the number of rules being reduced up to 21 times.

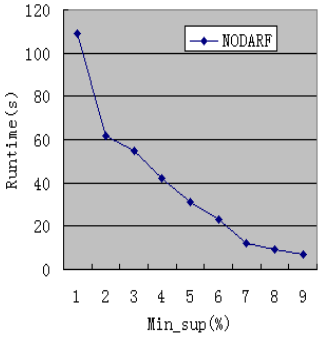

Fig 2. Runtime of NODARF

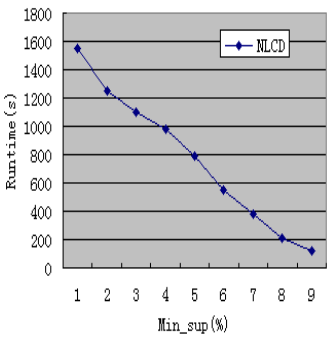

Fig 3. Runtime of NLCD

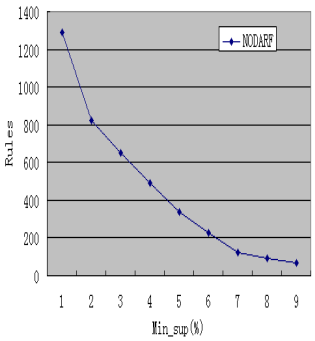

Fig 4. Rules of NODARF

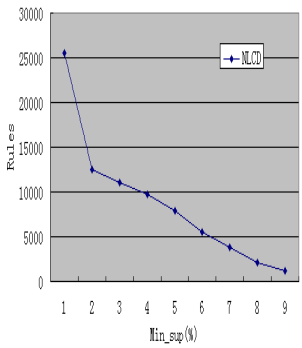

Fig 5. Rules of NLCD

5 Conclusions

The results of this study have shown how to use the application of non-redundant correlative rules mining technique in flow graph network for online instructors. From the experiment results it implies that this technique can be applied in order to obtain interesting information in an efficient and faster way. This approach can be improved on as following:

Firstly, instructors can use visualization techniques to obtain a general view of the student's usage data. For instance, if you find something strange or irregular in the plots, then you can obtain more detailed information about these events by viewing statistical values.

Secondly, for similar groups of students, you can apply clustering techniques in order to obtain the exact groups of students. And these groups can also be used to create a classifier in order to classify students. The classifier shows what the main characteristics of the students in each group are, and it allows new online students to be classified.

Finally, the instructors can apply association rule mining to discover if there is any relationship between these characteristics and other attributes. These rules can not only help to classify students, but also to detect the sources of any incongruous values obtained by the students.

Whereas the applying of this technique has not compared with other methods used in educational data mining field, the next work will develop a more efficient data mining tool based on non-redundant correlative rules mining technique with some distinguished methods.

Acknowledgment

This paper is sponsored by 2016 "Chuangxin Qiangxiao" Humanities & Social Sciences Project of the Education Department of Guangdong Province ,China(2016GXJK035), the National Natural Science Foundation of China(No.61370229), the S &T Projects of Guangdong Province, China (No. 2016 B010109008, 2015A030401087, 2015B010110002), and the S&T Projects of Guangzhou Municipality, China(No.201604010003, 201604010054).

References

[1] Agrawal, R.& Srikant, R. Fast algorithms for mining association rules [J]. In Pro of 20th International Conference on Very Large Data Bases, 1994, 487-499.

[2] Brin, S., Motwani, R., e.t. Dynamic itemset counting and implication rules for market basket data[J]. In Pro of ICMD, 1997, 255-264.

[3] Z. Pawlak, Flow graphs and data mining[J], Transactions on Rough Sets III, LNCS 3400, pp. 1–36, 2005.

[4] Bo Liu, Qi-lun Zheng, Hong Peng, Jin-song Hu, A Novel Algorithm of Holistic Correlation Analysis[J]. The Pro of ICMLC, 2005, 2145-2150.

[5] Bo Liu,Non-linear correlation discovery-based technique in data mining, 2008 International Symposium on Intelligent Information Technology Application (IITA 2008),pp:117-121,DEC. 21-22, Shanghai, China.

[6] Bo Liu,Jian-hua Zhao, Applying of non-linear correlation technology in educational data mining, 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery,14-16 August,Tianjing, China.

[7] David Lo, Siau-ChengKhoo, Limsoon Wong, Non-redundant sequential rules-Theory and algorithm, Information Systems 34(2009) 438-453.

[8] Bo Liu, Huijie Cui, Xiao Liu, and Changqin Huang. The Study on the Roles in the Practical Applying Oriented Asynchronous Collaborative Learning Network. ICBL 2017, Lecture Notes in Compter Science.(LNCS10309). pp.295-306, 2017-06.