김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제6호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 6 JUCHE106(2017).

망침입검출을 위한 드문련관규칙발굴의 한가지 방법

공혜옥, 정철영

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《수학, 물리학, 화학, 생물학과 같은 기초과학부문에서 과학기술발전의 원리적, 방법론적기초를 다져나가면서 세계적인 연구성과들을 내놓아야 합니다.》(《조선로동당 제7차대회에서 한 중앙위원회사업총화보고》 단행본 40페지)

론문에서 제기한 침입검출체계의 이상검출을 위한 드문련관규칙발굴알고리듬은 침입검출과 같이 드물게 발생하는 숨은 패턴들을 발굴하는데 적용할수 있으며 비빈발항목모임을 대상으로 하는것으로 하여 해쉬토대를 리용하는 전통적인 련관규칙발굴알고리듬보다 속도문제, 기억공간문제에서 우점을 가지고있다.

련관규칙발굴의 표준응용실례인 시장바구니분석의 목표는 업무자료기지안의 항목모임들사이의 관계를 밝히는것이다. 이 관계는 업무자료기지안의 빈발항목모임에 기초한다. 그러나 침입검출체계를 위한 망파케트자료기지에서는 시장바구니분석과는 달리 망가입자들의 정상활동이 비정상활동보다 많으므로 빈발항목모임을 발견하는것보다 비빈발항목모임에 기초하여 련관규칙발굴알고리듬의 효과성을 높일수 있다.

선행연구[1－3]에서는 공통적인 패턴과 규칙을 발굴하지만 고립적인 패턴이나 규칙 즉 드문련관규칙을 발굴하는데는 충분한 관심을 돌리지 못하였다.

선행연구[4]에서는 비빈발항목들로부터 효과적인 련관규칙을 발굴하는 알고리듬들을 제기하고 업무자료기지에서의 정량적련관규칙을 론의하였지만 관계형자료기지에서의 정량적련관규칙의 정식화를 주지 못하였으며 또한 정상적인것과의 차이를 표현하는 새로운 규칙을 이상규칙이라고 보고 드문패턴을 발굴하는 문제의 중요성을 제기하였으나 여러 형태의 이상패턴들을 효과적으로 식별하기 위한 흥미측도에 대하여서는 제기하지 못하였다.

우리는 침입검출에서와 같이 드물게 발생하지만 가치있는 패턴을 발굴하는것이 매우 중요하게 제기되는 응용분야들에 적용할수 있는 드문련관규칙발굴문제를 론의한다.

대부분의 련관규칙발굴알고리듬들은 흥미있는 규칙들을 발굴하기 위하여 최소지지도와 최소믿음도를 리용한다. 이때 이 2개의 파라메터에 의하여 생성되는 많은 련관규칙들이 축소되여도 여전히 사용자에게 흥미가 없는 많은 규칙들이 생성된다.

실례로 간장과 소금의 구입자료를 보여주는 표에서 규칙《간장→소금》의 지지도는 20%로서 높다. 이 규칙의 믿음도가 $p(\text{간장} \wedge \text{소금})/p(\text{간장}) = 20/25 = 0.8$ 즉 80%로서 상당히 높으므로 이 규칙은 유효한 규칙이라고 할수 있다. 그런데 소금을 사는 사람은 90%로서 믿음도 80%보다 크므로 이 규칙은 잘못 발굴되였다. 사실상 간장과 소금중에서 어느 하나를 구입하면 다른것을 구입하는 일은 거의 없으므로 서로 부정적으로 련관된다고 볼수 있다.

표. 간장과 소금의 구입자료(%)

	소금	ㄱ소금	Σ행
간장	20	5	25
ㄱ간장	70	5	75
Σ렬	90	10	100

우의 실례는 최소지지도와 최소믿음도를 리용하는 련관규칙발굴의 취약성을 보여준다. 즉 규칙에서 전제의 발생이 결론의 발생을 의미하지 않는다면 잘못된 련관규칙을 유도할 수 있다.

이로부터 론문에서는 호상관련성에 기초하여 자료항목모임들사이의 흥미있는 관계를 찾는 다른 방법을 제기한다.

주어진 자료기지안의 항목모임 A의 지지도를 $\mathrm{sprt}(A)$라고 할 때 자료기지안의 항목모임 X와 Y의 호상관련성에 대한 통계적정의는 다음과 같다.

$$\mathrm{Interest}(X,\ Y)=\mathrm{sprt}(X\cup Y)/(\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y))$$

이로부터 $\mathrm{Interest}(X,\ Y)=1$이면 $\mathrm{sprt}(X\cup Y)=\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)$이므로 Y와 X는 독립이며 $\mathrm{Interest}(X,\ Y)>1$이면 $\mathrm{sprt}(X\cup Y)>\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)$이므로 Y는 X에 긍정적으로 관련된다. 또한 $\mathrm{Interest}(X,\ Y)<1$이면 $\mathrm{sprt}(X\cup Y)<\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)$이므로 Y는 X에 부정적으로 관련된다. 즉 Interest값이 1이면 Y와 X는 독립이므로 규칙 $X\rightarrow Y$는 흥미가 없다. 이것은 전제와 결론이 독립이면 규칙이 흥미없다는것을 의미한다.

최소흥미도를 min_Interest라고 할 때 $|\mathrm{sprt}(X\cup Y)-\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)|\geq \mathrm{min_Interest}$이면 항목모임 $X\cup Y$는 잠정적으로 흥미있는 항목모임으로 볼수 있다.

$$\exists X,\ Y: X\cap Y=\phi,\ X\cup Y=Z\ ,\ \forall x_k\in X,\ y_k\in Y,$$
$$\mathrm{sprt}(x_k)\leq \mathrm{min\_sprt},\ \ \mathrm{sprt}(y_k)\leq \mathrm{min\_sprt}$$

일 때 Z는 잠정적으로 비빈발항목모임,

$$|\mathrm{sprt}(X\cup Y)-\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)|\geq \mathrm{min\_Interest}$$

이면 Z는 잠정적으로 흥미있는 비빈발항목모임이라고 한다.

min_sprt, min_conf, min_Interest를 각각 주어진 업무자료기지에서 규정된 최소지지도와 최소믿음도, 최소흥미도라고 할 때 다음과 같은 정의를 할수 있다.

정의 1 $I=\{I_1,\ I_2,\cdots,\ I_n\}$을 업무자료기지안의 항목들의 모임이라고 하고 $X,\ Y\subseteq I$, $X\cap Y=\phi$, $\mathrm{sprt}(X)\neq 0$, $\mathrm{sprt}(Y)\neq 0$, $\mathrm{min\_sprt}>0$, $\mathrm{min\_conf}>0$, $\mathrm{min_Interest}>0$일 때 다음의 조건 ①－③을 만족시키는 $X\rightarrow Y$를 흥미있는 유효규칙 혹은 련관규칙, 조건 ④－⑦을 만족시키는 $X\rightarrow Y$를 흥미있는 드문유효규칙 혹은 드문련관규칙이라고 한다.

① $\mathrm{sprt}(X\cup Y)\geq \mathrm{min\_sprt}$

② $|\mathrm{sprt}(X\cup Y)-\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)|\geq \mathrm{min\_Interest}$

③ $\mathrm{conf}(X\rightarrow Y)=\mathrm{sprt}(X\cup Y)/\mathrm{sprt}(X)\geq \mathrm{min\_conf}$

④ $\mathrm{sprt}(X)\leq \mathrm{min\_sprt},\ \mathrm{sprt}(Y)\leq \mathrm{min\_sprt}$

⑤ $|\mathrm{sprt}(X\cup Y)-\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)|\geq \mathrm{min\_Interest}$

⑥ $\mathrm{Interest}(X,\ Y)>1$

⑦ $\mathrm{conf}(X\rightarrow Y)=\mathrm{sprt}(X\cup Y)/\mathrm{sprt}(X)\geq \mathrm{min\_conf}$

R를 관계형자료기지, $f_1, f_2, \cdots, f_n$을 R의 마당들, q_{jk} $(j=1,\cdots,n,\ k=1,\cdots,j_m)$를 f_j의 리산화된 령역값들이라고 하면 R의 매 레코드 R_i는 유일한 식별자 ID를 가지며

$$R_i=\{f_1=q_{1k_i},\ f_2=v_{2k_i},\cdots,\ f_n=v_{nk_i},\ \ k_i=1,\cdots,j_m\}$$

으로 표시할수 있다. 따라서 R에서 가능한 모든 항목모임은 $I=\{R_i\,|\,i=1,2,\cdots,N\}$이다. 여기서 N은 R의 레코드수이다.

$X=\{f_1=p_1, f_2=p_2, \cdots, f_k=p_k\}\subseteq I$ 를 정량적항목모임이라고 할 때 R안의 어떤 레코드 RID에 대하여 $X\subseteq$RID이면 레코드 RID는 X를 포함한다고 말한다. 여기서 $f_1, \cdots, f_k$ 는 X안의 마당이름들이고 $p_1, \cdots, p_k$ 는 X안의 마당들에 대응되는 리산화된 령역값들이다.

관계형자료기지 R에서 정량적항목모임 X를 포함하는 레코드들의 백분률은 X의 지지도이다. 이것을 업무자료기지에서와 마찬가지로 $\mathrm{sprt}(X)$ 로 표시한다.

정의 2 $I=\{f_j=q_{jk}\mid j=1, 2, \cdots, n,\ k=1, \cdots, j_m\}$ 을 R안의 정량적항목모임이라고 하고

$$X,\ Y\subseteq I,\ X\cap Y=\phi,\ \mathrm{sprt}(X)\neq 0,\ \mathrm{sprt}(Y)\neq 0,$$
$$\mathrm{min\_sprt}>0,\ \mathrm{min\_conf}>0,\ \mathrm{min\_Interest}>0$$

일 때 다음의 조건 ①－③을 만족시키는 $X\rightarrow Y$ 를 정량적련관규칙, 조건 ④－⑦을 만족시키는 $X\rightarrow Y$ 를 드문정량적련관규칙이라고 한다.

① $\mathrm{sprt}(X\cup Y)\geq \mathrm{min\_sprt}$

② $|\mathrm{sprt}(X\cup Y)-\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)|\geq \mathrm{min\_Interest}$

③ $\mathrm{conf}(X\rightarrow Y)=\mathrm{sprt}(X\cup Y)/\mathrm{sprt}(X)\geq \mathrm{min\_conf}$

④ $\mathrm{sprt}(X)\leq \mathrm{min\_sprt},\ \mathrm{sprt}(Y)\leq \mathrm{min\_sprt}$

⑤ $|\mathrm{sprt}(X\cup Y)-\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)|\geq \mathrm{min\_Interest}$

⑥ $\mathrm{Interest}(X,\ Y)>1$

⑦ $\mathrm{conf}(X\rightarrow Y)=\mathrm{sprt}(X\cup Y)/\mathrm{sprt}(X)\geq \mathrm{min\_conf}$

드문정량적련관규칙의 발굴과정은 다음과 같은 두가지 단계들로 구성된다.

단계 1 흥미있는 비빈발항목모임들을 식별한다. 즉 정의 1(혹은 정의 2)의 ④로부터

$$\exists X, Y:\ X\cap Y=\phi,\ X\cup Y=Z,\ \forall x_k\in X,\ y_k\in Y,$$
$$\mathrm{sprt}(x_k)\leq \mathrm{min\_sprt},\ \ \mathrm{sprt}(y_k)\leq \mathrm{min\_sprt}$$

이면 Z는 비빈발항목모임이며 정의 2의 ⑤로부터

$$|\mathrm{sprt}(X\cup Y)-\mathrm{sprt}(X)\cdot\mathrm{sprt}(Y)|\geq \mathrm{min\_Interest}$$

이면 Z는 잠정적으로 흥미있는 비빈발항목모임이다. 이러한 Z들을 주어진 자료기지에서 모두 발견해낸다.

단계 2 발견된 비빈발항목모임들로부터 흥미있는 규칙들을 추출한다. 즉 비빈발항목모임들의 가능한 결합들로부터 정의 2의 ⑥을 리용하여 드문련관규칙들로 제한하며 정의 2의 ⑦을 리용하여 강한 흥미를 가진 드문련관규칙들을 추출한다.

이처럼 드문정량적련관규칙발굴과정은 련관규칙발굴과정과 원리적으로 동일하며 업무자료기지가 아니라 관계형자료기지라는것, 단계 1에서 빈발항목모임들이 아니라 비빈발항목모임들을 발견한다는것, 단계 2에서 흥미가 없는 련관규칙들은 배제한다는것 등에서 차이난다.

다음으로 망침입검출을 위한 해쉬화알고리듬에 대하여 보자.

망파케트자료기지는 관계형자료기지이며 침입자료는 정상자료보다 훨씬 드물게 나타난다. 따라서 망파케트자료기지에서 드문정량적련관규칙들을 발견하는 방법을 적용한 이상검출체계를 구축할수 있다.

또한 망파케트자료기지는 레코드길이가 상대적으로 길지 않고 동일하며 더우기 비빈발항목들의 수는 극히 적으므로 해쉬화방법을 적용하여 비빈발k－항목모임들의 지지도를 효

과적으로 계산할수 있다.

알고리듬은 다음과 같다.

입력: R, min_supp (최소지지도), min_Interest (최소흥미도), min_conf (최소믿음도)

출력: RAR (R의 드문련관규칙들의 모임)

① R를 조사하여 모든 1－비빈발항목모임들의 족 NL1을 구한다.

② Address $=\phi$

```
for ∀r∈R {
    rL1= 1-비빈발항목모임족(r); // rL1⊆NL1
    for(k=2; rLk-1≠φ; k++){
        rCk=apriori_gen(rLk-1);
        for ∀c∈rCk {
            Address(k, c) = Hash(c);
            if Address(k, c)∉Address{
                Address=Address∪Address(k, c);
                Value(k, c) = 1;}
            else
                Value(k, c)++;
        } //Hashing
    }
} //support
```

③ NL $=\phi$

```
for ∀Address(k, c)∈Address{
    NLk= {Hash-1(Address(k, c))|Value(k, c)≤min_supp};
    NL=NL∪NLk;
} //NL: All infrequent itemsets
```

④ RAR $=\phi$

$\forall X, Y \subseteq NL, X \cap Y=\phi$, if $(|supp(X \cup Y)-supp(X)supp(Y)| \geq min\_Interest) \wedge (Interest(X, Y)>1) \wedge (supp(X \cup Y)/supp(X) \geq min\_conf)$

```
RAR=RAR∪{X→Y}; //All Rare Association Rules
return RAR;

Procedure apriori_gen(NLk-1)
Ck=φ#
for ∀l1∈NLk-1{
    for ∀l2∈NLk-1{
      if (l1[1]=l2[1])∧(l1[2]=l2[2])∧…∧(l1[k-2]= l2[k-2])∧(l1[k-1]<l2[k-1]) then{
        c=l1∞l2;
        Ck=Ck∪k c;}}}
return Ck;
```

알고리듬에서는 먼저 자료기지를 조사하여 1－비빈발항목모임들을 식별하고 자료기지를 두번째로 조사하면서 매 레코드에 포함된 k－비빈발항목모임들의 지지도를 해쉬함수를 리용하여 계산한 다음

$$|\mathrm{sprt}(X \cup Y) - \mathrm{sprt}(X) \cdot \mathrm{sprt}(Y)| \geq \min\_\mathrm{Interest}\,,\ \ \mathrm{Interest}(X,\ Y) > 1$$

을 리용하여 흥미있는 정의 비빈발항목모임들만을 골라내며

$$\mathrm{conf}(X \to Y) \geq \min\_\mathrm{conf}$$

를 만족시키는 흥미가 강한 드문련관규칙을 발굴한다.

알고리듬의 적용과정은 아래와 같이 구성된다.

걸음 1 자료기지 R를 조사하면서 $\min_\mathrm{sprt} = 0.5$를 리용하여 1－비빈발항목모임 NL1을 구한다.

걸음 2 자료기지 R를 두번째로 조사하면서 매 레코드에 포함된 비빈발항목들을 식별해내고 그것들의 가능한 결합으로 k－비빈발항목모임들을 구한 다음 해쉬화수법으로 지지도를 계산한다.

걸음 3 $|\mathrm{sprt}(X \cup Y) - \mathrm{sprt}(X)\mathrm{sprt}(Y)|$의 값을 계산한다. 이 값이 min_Interest보다 작지 않은 흥미있는 드문련관규칙들을 추출한다.

걸음 4 $\mathrm{Interest}(X,\ Y) = \dfrac{\mathrm{sprt}(X \cup Y)}{\mathrm{sprt}(X) \cdot \mathrm{sprt}(Y)}$를 계산하고 $\mathrm{Interest}(X,\ Y) > 1$인 정의 련관규칙들을 추출한다.

걸음 5 $\mathrm{conf}(X \to Y) = \dfrac{\mathrm{sprt}(X \cup Y)}{\mathrm{sprt}(X)}$의 값을 계산하고 $\mathrm{conf}(X \to Y) \geq \min_\mathrm{conf}$인 흥미가 강한 련관규칙들을 추출한다.

이와 같이 우리는 마당개수가 수십개정도인 망파케트자료기지를 대상으로 하는 침입검출체계에서 드물게 발생하는 이상현상들을 목표로 침입패턴을 발굴할수 있는 실천적인 해쉬토대의 드문련관규칙발굴알고리듬을 개발하였다.

론문에서 제기한 알고리듬은 침입검출과 같이 드물게 발생하지만 가치있는 숨은 패턴을 발굴하는 응용분야들에 적용할수 있으며 비빈발항목모임을 대상으로 하기때문에 해쉬토대의 방법으로 전통적인 련관규칙발굴알고리듬보다 속도문제, 기억공간문제에서 뚜렷한 우점을 가지도록 설계되였다.

참 고 문 헌

[1] Hyeok Kong et al.; IJTPC, **10**, **12**, 1, 2015.

[2] K. M. M. Aung et al.; Proceedings of 2015 International Conference on Future Computational Technologies, 29～30, 164～170, 2015.

[3] Li Hanguang et al.; Physics Procedia, **24**, 1615, 2012.

[4] N. Rountree et al.; Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection, Hershey, 15～32, 168～184, 2010.

주체106(2017)년 2월 5일 원고접수

A Method of Rare Association Rule Mining for Network Intrusion Detection

Kong Hye Ok, Jong Chol Yong

We propose a new practical rare association rule mining algorithm for anomaly detection in intrusion detection system (IDS). This algorithm can be applied to the fields require to mine the hidden patterns which are rare but valuable like IDS, and it is designed based on hashing among infrequent item-sets, and it has obvious advantages of speed and memory space limitation problems than the traditional association rule mining algorithms.

Key words: infrequent item-set, anomaly detection