Mining h-Dimensional Enhanced Semantic Association Rule Based on Immune-Based Gene Expression Programming*

Tao Zeng[1], Changjie Tang[1], Yintian Liu[1], Jiangtao Qiu[1], Mingfang Zhu[1,2]
Shucheng Dai[1], and Yong Xiang[1,3]

[1] School of Computer, Sichuan Univ., Chengdu, 610065, China
{zengtao, tangchangjie}@cs.scu.edu.cn
[2] Dept. of Computer Sci. & Tech., Shaanxi Univ. of Tech., Hanzhong, 723003 China
[3] Chengdu Electromechanical college, Chengdu, 610031, China

Abstract. Rule mining is very important for data mining. However, traditional association rule is relatively weak in semantic representation. To address it, the main contributions of this paper included: (1) proposing formal concepts on ***h***-**D**imensional **E**nhanced **S**emantic **A**ssociation **R**ule (*h*-DESAR) with self-contained logic operator; (2) proposing the *h*-D**E**SA**R** mining method based on **I**mmune-based **G**ene Expression Programming (ERIG); (3) presenting some novel key techniques in ERIG. Experimental results showed that ERIG is feasible, effective and stable.

1 Introduction

Rule mining is an important task of data mining because it is easy to understand rules better than other data mining model. Association rule (AR) mining has been a hot research theme in data mining due to its broad applications at mining association, correlation, causality, and many other important data mining tasks [1-4].Fruitful research results for AR mining can be found in [1-4].

However, complex data mining application requires refined and rich-semantic knowledge representation. Traditional association rule is relatively weak in semantic representation. Example 1 and 2 show that it is difficult for traditional concepts and methods to describe and discover rich-semantic rule.

Example 1. Customers probably purchase "laptop" if age is "30-40", title is "*prof.*", and address is ***not*** at "campus". To describe this fact, we need new association rule in the form of

$$age("30-40") \wedge title("prof.") \wedge (\neg address("campus")) \rightarrow purchase("laptop") \quad (1)$$

Example 2. Customers probably purchase "PC" if age is "30-40", ***either*** title is "ass.", ***or*** address is at "campus". To describe this fact, we need other new association rule in the form of

* This paper was supported by the National Science Foundation of China under Grant Nos. 60473071 and 90409007.

L. Feng et al. (Eds.): WISE 2006 Workshops, LNCS 4256, pp. 49 – 60, 2006.

$$age("30-40") \wedge (title("ass.") \vee address("campus")) \rightarrow purchase("PC") \quad (2)$$

On issues like Example 1 and 2, we can retrieve little related work except [5]. In 2002, Zuo proposed an effective approach based on **G**ene **E**xpression **P**rogramming (GEP) to mine **P**redicate **A**ssociation **R**ule (PAR), named PAGEP [5].

However, PAGEP's main objective to mine is single-dimensional. And PAGEP can not always success in discovering strong PARs stably.

To address it, focusing on multi-dimensional problem, we proposed algorithms based on Immune-based GEP to mine ***h*-D**imensional **E**nhanced **S**emantic **A**ssociation **R**ule (*h*-DESAR). These are distinguished from [5] and other related works.

The main contributions of this work included that formal concepts and properties of *h*-DESAR were proposed, and the *h*-D**E**SA**R** mining algorithm based on **I**mmune-based **G**EP (ERIG) was proposed, implemented and tested.

Main novel techniques in our ERIG include:

- The distinctive structures of immune cell and antibody, which can carry 8 pieces of *h*-DESARs to decrease computing complexity 8 times;
- The Dynamic Self-Tolerance Strategy where self set can change dynamically and both invalid and redundant immune cell can be eliminated.
- The heuristic *h*-D**E**SARs **R**eduction **C**riterion (EPC), that is, a strong rule is fine if and only if the contra-positive of it is strong too.

The remaining of the paper is organized as follows. Section 2 describes the background of problem and our motivation. Section 3 formally introduces the problem. Section 4 proposes the ERIG algorithm, presents some distinctive methods or strategies and discusses the time complexity. Section 5 gives experimental results. Finally, Section 6 is conclusion and future work.

2 Background and Motivation

2.1 Background

Gene Expression Programming (GEP) [5,6,7] is of genetic computing introduced by Candida in 2001 [6]. The philosophical strategy hidden in GEP is to solve complex problem with simple code. GEP is somewhat similar to, but not the same as Genetic

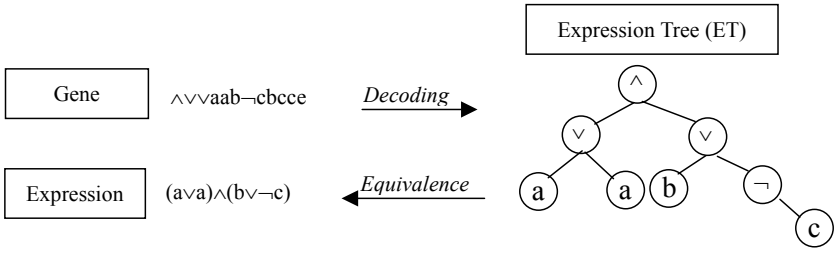

Fig. 1. Decoding process in GEP

Algorithms (GA) or Genetic Programming (GP). The chromosome of GP is tree-formed structure directly, while that of GEP is linear string. So GP's genetic operations are designed to manipulate the tree forms of chromosomes. However, GEP's genetic operations are similar to but simpler than those in GA. Compared with its ancestors, GEP innovated in structure and method. It uses a very smart method to decode gene to a formula [5,6,7]. Fig. 1 demonstrates the decoding process in GEP.

As an example, if let "a", "b" and "c" represent atomic predicates "*age*(*x*)", "*title*(*x*)" and "*address*(*x*)" respectively, then the expression in Fig. 1 can express the logic formula "$(age(x) \vee age(x)) \wedge (tile(x) \vee \neg address(x))$".

Like the example above, this paper will utilize GEP to express and discover the predicate formulas that can be used to construct enhanced semantic meta-rule. Please refer to [5, 6, 7] for the other detailed description on GEP due to the limited space.

Artificial Immune System (AIS) [9-12] is a rapidly growing field of information processing based on immune inspired paradigms of nonlinear dynamics. It is expected that AIS, based on immunological principles, be good at modularity, autonomy, redundancy, adaptability, distribution, diversity and so on. As a member of nature-inspired computing, AIS imitates biology immune system, aiming not only at a better understanding of the system, but also at solving engineering problems.

There are various models or techniques for AIS based on different algorithms or representations. According to [10], the main representations used include binary strings, real-valued vectors, strings from a finite alphabet, java objects and so on.

2.2 Motivation of Proposing Immune-Based GEP

GEP is strong in representing and discovering knowledge with simply linear strings. AIS has many advantages in evolution control. It is natural to assume that embedding GEP in AIS will inherit and enhance advantages of AIS and GEP.

3 Formal Statements for *h*-DESAR

This section introduces some notations, presents the formal statement of problems, and discusses their properties. Basic relational algebra notations come from [8].

Let S^m denote a *m*-dimensional relation instance, $Attr(S^m)$ denote attribute symbol set of S^m, $Dom(A_i)$ denote the domain of attribute A_i, $\hat{S}^m = (A_1, A_2, ..., A_m)$ denote relation schema, and $t = (V_{A_1}, V_{A_2}, ..., V_{A_m})$ denote a tuple of S^m, where $A_i \subseteq Attr(S^m)$, $V_{A_i} \in Dom(A_i)$ for i=1,..,m.

Let H = {y | y is a well-formed predicate formula}, $Ary(h)$ be arity of h, and $PreSymbol(h)$ = {z | z is the symbol of an atomic formula in h, $h \in H$ }. For instance, if q is $A(x) \vee A(x) \wedge (B(x) \vee \neg C(x))$, then $Ary(q)$ = 3 and $PreSymbol(q)$ = {A, B, C}. Let W be a set, $|W|$ denote the size of W that is the number of elements in W, and $\#(S^m)$ denote record number of S^m.

3.1 Enhanced Semantic Meta-rule

The following formal statements on enhanced semantic meta-rule are different from those proposed by Fu in [4].

Definition 1. An ***enhaced semantic meta-rule*** $\mathfrak{R}$ on S^m can be described as a logic formula in the form of $P \rightarrow Q$, where

- Let $X, Y \subset Atrr(S^m)$, $X \neq \phi$, $Y \neq \phi$, $X \cap Y = \phi$, and $\Omega_M = \{\psi \mid \psi$ is an atomic first-order predicate whose symbol is in M, and $\psi(x)$ means the value of attribute ψ is $x\}$ and $F = \{\neg, \wedge, \vee\}$.
- P is a well-formed first-order logic formula composed of the atomic formulas in Ω_X and logic operators in F.
- Q is a well-formed first-order logic formula composed of the atomic formulas in Ω_Y and logic operators in F.

Additionally, we call P **antecedent**, Q **consequent** and $\{P, Q\}$ **foundation set** of it. □

Definition 2. We call an enhanced semantic meta-rule ***h-dimensional enhanced semantic meta-rule*** $\mathfrak{R}_h$ if and only if

- $2 \leq h \leq m$, and $ary(P) + ary(Q) = h$.
- P and Q have been *simplified*.[1]
- The atomic predicates in P and Q can occur *only once* in P and Q respectively.

Additionally, let $\{\mathfrak{R}_h\}$ denote the set of all h-dimensional enhanced semantic meta-rules on S^m. □

Remark 1. The logic operators we used include "AND", "OR", "NOT" which is self-contained.

3.2 *h*-Dimensional Enhanced Semantic Association Rule

Let M be an attribute set, $M \subseteq Attr(S^m)$ and the sub-tuple $GetFTuple\ (M,\ t) = \prod_{A_1^M, A_2^M, \ldots, A_{|M|}^M} t$ where $A_i^M \in M$ for $i=1,\ldots,|M|$.

Definition 3. Given a tuple $t = (V_{A_1}, V_{A_2}, \ldots, V_{A_n}) \in S^m$ and $\mathfrak{R}_h$, a ***h-Dimensional Enhanced Semantic Association Rule*** $\mathfrak{R}_{h,t}^s$ can be described a logic formula in the form of $P^s \rightarrow Q^s$, where

- Let P be antecedent and Q be consequent of $\mathfrak{R}_h$.
- P^s is the substitution formula of P, in which all variables were replaced by the corresponding value in $GetFTuple\ (P,\ t)$ according to the meanings of atomic predicate in P.
- Q^s is the substitution formula of Q, in which all variables were replaced by the corresponding value in $GetFTuple\ (Q,\ t)$ according to the meanings of atomic predicate in Q.

Additionally, we call the tuple t ***feature tuple***. P^s and Q^s is antecedent and consequent of it respectively. $\{ P^s, Q^s \}$ is foundation set of it. □

[1] Here *simplified* means that expression string is parsed to create an expression tree and both redundant brackets and "¬" are eliminated from expression string. For example, "(((a)∧b))" can be simplified to "a∧b", "¬¬¬a" to "¬a", and "¬¬¬¬a" to "a".

It is obvious that both Example 1 and Example 2 are sound 4-dimensional enhanced semantic association rules.

Given two *m*-dimensional tuples $t_1=(V_{c_1},V_{c_2},...,V_{c_m})$ and $t_2=(V_{c_1}',V_{c_2}',...,V_{c_m}')$, let $\bar{t}=(f_{c_1},f_{c_2},...,f_{c_m})$ denote ***match tuple*** between t_1 and t_2 where

$$f_{C_j}=\begin{cases}\text{true} & \text{if } V_{C_j}=V_{C_j}'\\ \text{false} & \text{if } V_{C_j}\neq V_{C_j}'\end{cases}\qquad j=1,\ldots,m \tag{3}$$

Definition 4. Let U be one of P^s, Q^s and $P^s\wedge Q^s$ of $\Re_{h,t}^{\ s}$, and t be feature tuple of $\Re_{h,t}^{\ s}$. For $\forall t'\in S^m$, we say that t' ***support*** U if and only if

- Let $\bar{t}$ be the match tuple between t and t', and $\bar{t}_P=GetFTuple(Attr(P^S{}_x),\ \bar{t}\ )$.
- U^S is the boolean formula substituted for U, in which all atomic predicates were replaced by the corresponding boolean value in $\bar{t}_P$ according to the mapping relationship between attributes in $\bar{t}_P$ and atomic predicates in U.
- Evaluate U^S and the result is *true*.
- Otherwise, t' does ***not support*** U. □

Let $u\in\{P^s, Q^s, P^s\wedge Q^s\}$, and $\sigma(u \mid S^m)$ denote the number of records that support u. The ***support degree*** and ***confidence degree*** can be described as follows.

- **Support degree:** $$sup(\Re_{h,t}^{\ s}, S^m)=\frac{\sigma(P^S\wedge Q^S\mid S^m)}{\#(S^m)} \tag{4}$$

- **Confidence degree:** $$conf(\Re_{h,t}^{\ s}, S^m)=\frac{\sigma(P^S\wedge Q^S\mid S^m)}{\sigma(P^S\mid S^m)} \tag{5}$$

Let min_conf, $min\_sup\in[0, 1]$. $\Re_{h,t}^{\ s}$ is ***strong*** if and only if $sup(\Re_{h,t}^{\ s}, S^m)\geq min\_sup$ and $conf(\Re_{h,t}^{\ s}, S^m)\geq min\_conf$ like [1,3].

3.3 Example

Example 3. 1) Let $F_A = age(x)\wedge(title(x)\vee\neg address(x))$, $F_A' = (age(x)\vee age(x))\wedge(title(x)\vee\neg address(x))$ and $F_B = purchase(x)$ where "age, title, address, purchase"$\in Attr(S^m)$. Then both $F_A\rightarrow F_B$ and $F_A'\rightarrow F_B$ are well-formed enhanced semantic meta-rule, but only $F_A\rightarrow F_B$ comply with *h*-dimensional enhanced semantic meta-rule according to Definition 2 where $h = 4$.

2) Given a tuple r("30", "male", "campus", "*prof.*.", "laptop")$\in S^m$ and $\hat{S}^m$ = (age, gender, address, title, purchase), $F_A{}^S\rightarrow F_B{}^S$ is a $\mathfrak{R}_{4,r}^{\ s}$ where $F_A{}^S = age$("30") $\wedge$ ($title$("*prof.*") $\vee\neg address$("campus")) and $F_B{}^S = purchase$("laptop").

3) Suppose that there is another tuple r'("30", "male", "not in campus", "*ass.*.", "laptop") $\in S^m$, then match tuple $\bar{r}$ between r and r' is (true, true, false, false, true). Because $u^S{}_A$ = true$\wedge$(false$\vee\neg$ false) = true, $u^S{}_B$ = true and $u^S{}_A\wedge u^S{}_B$ = true, r' support $F_A{}^S$, $F_B{}^S$ and $F_A{}^S\wedge F_B{}^S$. □

In this paper, we focus on mining *h*-DESAR, in which the atomic predicates in it occur only once, because it is more extractive and heuristic.

3.4 Some Properties of *h*-DESAR

Lemma 1. If *FS* is a foundation set, then *FS* can be used to construct 8 pieces of *h*-DESARs. They can be grouped into 4 pairs. Two *h*-DESARs in each pair are equivalent in logic each other.

Proof. Suppose that there is a foundation set $FS = \{A, B\}$, and we can construct the following 8 *h*-DESARs: 1) $A \rightarrow B$, 2) $\neg B \rightarrow \neg A$, 3) $B \rightarrow A$, 4) $\neg A \rightarrow \neg B$, 5) $\neg A \rightarrow B$, 6) $\neg B \rightarrow A$, 7) $A \rightarrow \neg B$, and 8) $B \rightarrow \neg A$. In them, 1) and 2), 3) and 4), 5) and 6), 7) and 8) are the contra-positive each other respectively. Since the contra-positive is equivalent to the original statement, two statements in pair are equivalent each other. □

Theorem 1. Let $FS=\{A, B\}$ be a foundation set and S^m be a relation instance. If $\sigma(A|S^m)$, $\sigma(B|S^m)$, $\sigma(A \wedge B|S^m)$ and $\#(S^m)$ were given, then all of support degree and confidence degree for 8 pieces of *h*-DESARs constructed by *FS* can be evaluated.

Proof. Because in system, arbitrary tuple can either support *h*-DESAR or not, we can compute the following value: 1) $\sigma(\neg A|S^m) = \#(S^m) - \sigma(A|S^m)$, 2) $\sigma(\neg B|S^m) = \#(S^m) - \sigma(B|S^m)$, 3) $\sigma(A \wedge (\neg B)|S^m) = \sigma(A|S^m) - \sigma(A \wedge B|S^m)$, 4) $\sigma(\neg A \wedge B|S^m) = \sigma(B|S^m) - \sigma(A \wedge B|S^m)$, 5) $\sigma(\neg A \wedge \neg B|S^m) = \#(S^m) - \sigma(A|S^m) - \sigma(B|S^m) + \sigma(A \wedge B|S^m)$. we can use these values to evaluate all support degree and confidence degree for these *h*-DESARs according to Equation (4) and (5). □

Lemma 2. Given a relation instance S^m and an enhanced semantic meta-rule $\mathfrak{R}_h$, let *EARSet* be the set of enhanced semantic association rule complied with $\mathfrak{R}_h$ on S^m, then $|EARSet| \leq \#(S^m)$.

Proof. According to definition 3, let $W = PreSymbol(P^S) \cup PreSymbol(Q^S)$, a sub-tuple *GetFTuple* (W, t) can be corresponding to a *h*-DESAR. Two cases arise: (a) If each of such sub-tuple in S^m is unique, then $|EARSet| = \#(S^m)$. (b) If there exist any duplicate sub-tuples in S^m, then $|EARSet| < \#(S^m)$. So $|EARSet| \leq \#(S^m)$. □

4 The ERIG Algorithm

4.1 Framework

We call our algorithm ERIG (the *h*-D**E**SA**R** mining based on **I**mmune-based **G**ene Expression Programming). The AIS in ERIG is somewhat similar to the hybrid of the clonal selection principle [9-10] and the negative selection algorithm [11]. However, different from other models, the representation in our AIS is gene of GEP and mutation operators come from GEP. In addition, many new techniques were proposed in ERIG. The algorithm framework is as follows.

Algorithm 1. (ERIG) The *h-Dimensional **E**nhanced Semantic Association **R**ule mining based on **I**mmune-based **G**ene Expression Programming.*
Input: *A m-dimensional relation instance S^m, a minimum support, minsup, and a minimum confidence, minconf.*
Output: *The set of strong h-Dimensional Enhanced Semantic Association Rules.*

```
BEGIN
1  Initialize and set control parameters;
   // The cellnum is the number of cells every generation. A outer loop is a generation.
   // The hfmt is the high frequent mutation threshold
2   WHILE stop condition is not satisfied BEGIN
3     BCSet := NULL; // BCSet is immune cells set
4     count := 0;
5     WHILE BCSet.size < cellnum AND count < hfmt BEGIN
6       BCSet := GenBCSet(cellnum, Ŝ^m, F, control parameters);
        //Call GenBCSet to generate BCSet via GEP
7       BCSet := SelfTolerance(BCSet);     // Self tolerance
8       count ++ ;
9     END WHILE
10    ABSet := MaturateCells(BCSet, S^m);  // Produce antibody set ABSet
11    Maturateaffinity(ABSet, BCSet, minsup, minconf);
      // Evaluate and eliminate those cells and antibodies which can not meet requirement.
12    MemorizeCells(BCSet);     // Add cells in BCSet to elite gene pool for GEP
13    Output(ABSet);            // Output solution for problem;
14    CloneMutation(BCSet);
15  END WHILE
END.                                                                    □
```

The code is self-explanatory. But it is impossible to list all detail of ERIG. We will select some distinctive methods or strategies to show as follows.

4.2 Some Key Techniques in ERIG

4.2.1 Structures of Immune Cell and Antibody

Immune cell and antibody are very important for AIS. In general, antigen is corresponding to the problem to be solved and antibody to the solution for it. For *h*-DESAR problem, the record in relation instance can be antigen and *h*-DESAR can be antibody. Through comprehensive analysis on each aspect, we designed our antibody and immune cell (B cell). The formal definition is as follows.

Definition 5. An immune cell, ***BCell***, is a 3-tuple (G, E, δ) where

- $G = (g_A, g_B)$, called ***chromosome***, is a 2-tuple, where both g_A and g_B are genes of GEP.
- $E = (e_A, e_B)$, called ***dual-expression***, is a 2-tuple, which were decoded from genes in G according to GEP.
- $\delta \in \{-1, 0, 1, 2\}$ is the state value of *BCell*, where –1, 0, 1 and 2 indicate cell is dead, immature, mature and memorized respectively. □

Definition 6. An ***antibody*** is a 3-tuple, (E, L, V), where

- E comes from the immune cell that produces this antibody.
- $L = (l_A, l_B)$ is a 2-tuple, where l_A and l_B are the substitution formulas for those in E respectively by attribute values of record in relation instance.
- The 4-tuple $V = (p_A, p_B, p_{AB}, p_{total})$ stores information about affinity where p_A, p_B, p_{AB} and p_{total} are the support number of l_A, l_B and $l_A \wedge l_B$ and the total number of records who were tested respectively. □

In Table 1 and Table 2, examples for *BCell* and *Antibody* were given respectively.

Table 1. An example for *BCell*

No.	Symbol	Value	Symbol	Value
1	$G.g_A$	∧∨∨aab¬cbcce	$G.g_B$	da∧e∨a…
2	$E.e_A$	(a∨a)∧(b∨¬c)	$E.e_B$	d
3	δ	0		

Table 2. An example for *Antibody*

No.	Symbol	Value	Symbol	Value
1	$E.e_A$	(a∨a)∧(b∨¬c)	$E.e_B$	d
2	$L.l_A$	(*age*(“30”) ∨ *age*(“30”)) ∧ (*title*(“*ass.*”) ∨ ¬ *address*(“campus”))	$L.l_B$	*purchase*(“PC”)
3	$V.p_A$	60	$V.p_B$	87
4	$V.p_{AB}$	58	$V.p_{total}$	100

Theorem 2. An antibody can represent and evaluate 8 pieces of *h*-DESARs.

Proof. Let *Ab* denote an antibody and by Lemma 1, use $\{Ab.L.l_A, Ab.L.l_B\}$ to construct 8 pieces of *h*-DESARs. Then, after affinity maturation, there are $\sigma(Ab.L.l_A|S^m) = Ab.V.p_A$, $\sigma(Ab.L.l_B|S^m) = Ab.V.p_B$, $\sigma(Ab.L.l_A \wedge Ab.L.l_B|S^m) = Ab.V.p_{AB}$, and $\#(S^m) = Ab.V.p_{total}$. Thus we can evaluate these 8 *h*-DESARs by Theorem 1. □

It shows our antibody is good at representation and discovery of *h*-DESARs.

4.2.2 Dynamic Self-tolerance Strategy

The part of self-tolerance in ERIG develops from negative select algorithm [11] and looks like that in [12]. However there are many differences among them. Our self-tolerance strategy is problem-oriented. Main strategy is as follows.

- Treat those immune cells that have been generated or used as self dynamically.
- Let *Bc* be an immune cell, and *SS* be self-set. For $\forall Bc \in SS$ where $Bc.E=(e_A,e_B)$, those cells are self too, if their dual-expression is one of (e_B, e_A), $(\neg e_A, e_B)$, $(e_B, \neg e_A)$, $(e_A, \neg e_B)$, $(\neg e_B, e_A)$, $(\neg e_A, \neg e_B)$ and $(\neg e_B, \neg e_A)$.
- Inject vaccine if it is needed. And treat those cells with certain pattern as self.

The function of our self-tolerance strategy is as follows.

- Avoid generating redundant cells that are equivalent to represent *h*-DESARs.
- Avoid generating any fault cells that cannot represent valid *h*-DESARs.
- Be able to inject vaccine too.

4.2.3 Affinity Computing

In course of *affinity maturation*, for each antibody, its affinity information for all records (antigens) will be computed. After affinity maturation, there are $\sigma(Ab.L.l_A|S^m) = Ab.V.p_A$, $\sigma(Ab.L.l_B|S^m) = Ab.V.p_B$, $\sigma(Ab.L.l_A \wedge Ab.L.l_B|S^m) = Ab.V.p_{AB}$, and $\#(S^m) = Ab.V.p_{total}$. According to Theorem 1 and Theorem 2, we can scan database once but evaluate 8 times more *h*-DESARs than antibodies.

Additionally, because the statement and contra-positive is logically equivalent, we proposed a heuristic *h*-DESARs **R**eduction **C**riterion (EPC) to reduce result set, that is, a strong rule is fine if and only if the contra-positive of that is strong.

4.3 Algorithms Analysis

In this section, we discuss the time complexity of ERIG.

Theorem 3. Let h be a constant and $\#(S^m) = n$, then the time complexity of each generation in ERIG depends on the number of antibodies, and it is lower than $O(n^2)$.

Proof. Since operation on database is time-consuming, $\#(S^m)$ is the variable that has great impact on the time complexity. The number of cells, the high frequent mutation threshold, the size of pool and other control parameter all are limited constants. Hence, the time complexity from row 2 to 9 in ERIG is bounded above. It is $O(C_1)$. Similarly, *MemorizeCells* and *CloneMutation* are $O(C_2)$ and $O(C_3)$. Supposing the maximum number of cells of every generation is c, then the time complexity of *MaturateCells* is lower than $O(c*n^2)$. It is because, under worst-case condition, c cells can produce $c*n$ antibodies with scanning database once and each of these $c*n$ antibodies will match with n tuples to compute affinity. Finally, for *Output*, there are $c*n$ antibodies to process at most. So it is $O(c*n)$. To sum up, the total maximum time complexity is $O(C_1)+O(C_2)+O(C_3)+O(c*n^2)+O(c*n) \approx O(n^2)$. □

5 Experiments

To verify our algorithm, various cases were designed. The test platform is as follows: CPU: Intel C3 1.0GHz, memory: 320MB, hard disk: 80GB, OS: MS Windows XP Pro. SP2, compiler: JDK1.5.03. The data set we used in our experiments is *cmc*, with 10 dimensions and 1473 rows. It comes from UCI Machine Learning Repository[2]. Table 3 gives us symbol definitions for this section.

5.1 Case Test

Because there has been little research on *h*-DESAR, the case 1 was designed to compare ERIG with traditional AR mining.

Case 1. Let $F = \{\wedge\}$, *minsup*=0.5%, *minconf*=95%, *cellnum* = 20, and *hfmt* = 200. We run ERIG and Apriori algorithm to mine traditional multi-dimensional AR on data set "*cmc*" respectively to verify ERIG.

Remark 2. If $F = \{\wedge\}$ and the order of predicates not be considered, *h*-DESAR is equivalent to traditional multi-dimensional AR.

In this case, in order to utilize Apriori algorithm to mine multi-dimensional AR, we preprocess *cmc* in the following way. For each value of attribute in *cmc*, we add a string of its attribute in front of it to construct a new value, whose type become string, then store it into a new data set *cmc'*. After preprocessing, in *cmc'*, original equal values of different dimensions in *cmc* became unequal. It will eliminate possible value-collision between dimensions when Apriori runs on *cmc'*.

[2] http://www.ics.uci.edu/~mlearn/MLRepository.html

Some details of the result for case 1 were gives in Table 4. It showed the number and content of ARs mined by ERIG on *cmc* are the same as those by Apriori on *cmc'*.

Table 3. Symbol Definitions for Section 5

Symbol	Definition
TC	Total number of independent cells
TSAR	Total number of strong *h*-DESARs
EGN	The generation number when program ends
ACGE	Percentage of cell generation efficiency where ACGE = TC/(EGN**cellnum*) * 100 %
ECN	Number of cells eliminated by self tolerance
AREC	Ratio of ECN and TC where AREC = ECN / TC (times)
CEE	Number of cells generation by elite pool
AEE	Percentage of usage of elite pool where AEE = CEE / TC * 100%

Table 4. Results for Case 1 minsup=0.5%, minconf=95%

No.	*h*	EGN	TC	TSAR	ACGE	AREC	AEE
1	2	3	45	35	75.0%	2289	8.9%
2	3	19	360	744	94.7%	291	30.6%
3	4	75	1470	3418	98.0%	34	25.6%
4	5	197	3780	6324	95.9%	23	21.0%
5	6	336	6510	5650	96.9%	18	13.7%
6	7	391	7560	2470	96.7%	25	5.1%
7	8	295	5715	478	96.9%	30	0.7%
8	9	132	2550	22	96.6%	52	0.0%
9	10	28	511	0	91.3%	221	0.0%

Note: Sum of TSAR is 19141. The number and content are the same as those via Apriori on cmc'.

Table 5. Results for Case 2 *minsup* = 5% *minconf* = 98.5%

No.	*h*	EPC	Convergence	Vaccine	EGN	TC	TSAR	ACGE	AREC	AEE
1	3	no	yes	no	292	5760	316292	98.6%	19.26	50.0%
2	3	yes	yes	no	297	5760	980	97.0%	24.03	20.9%
3	4	no	no	no	500	10000	1334128	100.0%	0.00	49.8%
4	4	yes	no	no	500	10000	6796	100.0%	0.01	30.6%
5	5	yes	no	no	500	10000	18552	100.0%	0.00	30.6%
6	7	yes	no	no	500	10000	36984	100.0%	0.00	16.6%
7	2 to 6	yes	no	yes	500	10000	6431	100.0%	0.09	14.1%
8	7	yes	no	yes	500	10000	5434	100.0%	0.01	1.7%

Notes – In No. 7, the dimensions were restricted to 2nd, 3rd, 4th, 6th, 7th and 8th.
– In No. 8, the dual-expression template is ("#", "(#∨#)∧ (#∨¬#)∧(¬#∨#)").

Case 2. Let $F = \{\neg, \wedge, \vee\}$, *cellnum* = 20, *hfmt* = 200, and the order of predicates be considered. On data set "*cmc*", mine general *h*-DESARs, restricted dimensional

h-DESARs and the special *h*-DESARs generated by fixed dual-expression template, which test the function of vaccine respectively.

The results for Case 2 are in Table 5. Extensional tests show that 1) our algorithm is stable, 2) the efficiency of EPC is notable by comparison, 3) the capability of generating new immune cells is strong, and 4) the function of vaccine is sound and effective. As an example, a 5-DESAR from results of No.7 in Tab 5 is as follows.

$$D_8(1) \rightarrow D_7(3) \vee \neg (D_3(4) \wedge D_4(1) \wedge D_2(4)) \quad \text{sup} = 8.7\% \quad \text{conf} = 99.22\% \quad (6)$$

$$\neg(D_7(3) \vee \neg (D_3(4) \wedge D_4(1) \wedge D_2(4))) \rightarrow \neg D_8(1) \quad \text{sup} = 6.5\% \quad \text{conf} = 98.97\% \quad (7)$$

Since the 5-DESAR (6) and (7) are equivalent each other, they can be reduce to a 5-DESAR, where D_i denote *i*th dimension.

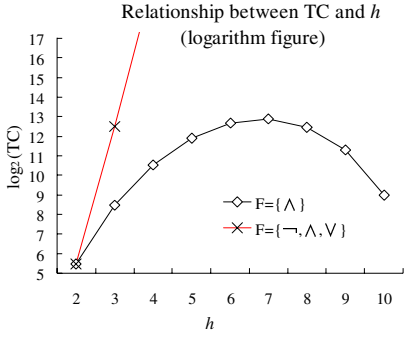

Fig. 2. Relationship between TC and *h* in Case 1 and Case 2

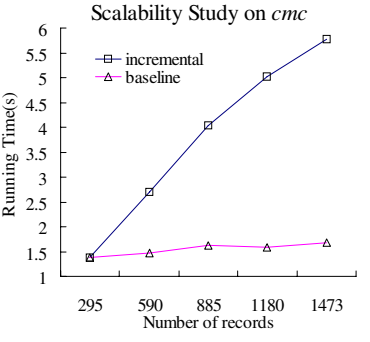

Fig. 3. Relationship between average running time per generation and record number of data set in Case 3

5.2 Scalability Study

In this section, we study the impact of relation instance scale on the performance of our algorithm.

Case 3. Let $F = \{\neg, \wedge, \vee\}$, *cellnum* = 20, *hfmt* = 200, $h = 4$ and do the following.

- Take a copy of records from row 1 to 295 in *cmc* as a new data set cmc_1, similarly, records from row 1 to 590 as cmc_2, records from row 1 to 885 as cmc_3, and records from row 1 to 1180 as cmc_4;
- Merge 2 copies of cmc_1 into a new data set cmc_2', similarly, 3 copies of cmc_1 into cmc_3', 4 copies of cmc_1 into cmc_4', and 5 copies of cmc_1 into cmc_5';
- Mine *h*-DESARs respectively on cmc_1, cmc_2, cmc_3, cmc_4, and *cmc* up to 100 generations several times.
- Similarly, do it respectively on cmc_1, cmc_2', cmc_3', cmc_4', and cmc_5' as baseline.

Fig. 3 described the result. The number of distinct tuples from cmc_2', cmc_3', cmc_4' to cmc_5'are the same as those in cmc_1, besides different in the number of records, so that the number of antibodies does not change and average running time per generation increases very slowly in baseline. However, when we do step 3) in Case 3, the number of unique tuples from cmc_1, cmc_2, cmc_3, cmc_4 to *cmc* increases gradually

with the rise of rows. Thus for every generation, more antibodies are generated and the running time ascends. But it is not so steep. It testifies to Theorem 3.

In Table 6, we bring a comparison between ERIG, PAGEP, and Apriori on available objective to mine.

Table 6. Comparison between ERIG, PAGEP, and Apriori

Available objective to mine	ERIG	PAGEP	Apriori
Traditional association rule	✓	✓	✓
Rule with connectives beyond "∧"	✓	✓	×
Rule with constrained pattern	✓	×	×
Rule with constrained attributes	✓	×	×

6 Conclusion and Future Work

We have discussed *h*-DESAR problem, proposed ERIG algorithm, presented some key techniques in ERIG. Experimental results testified to our expectations and showed that the ERIG is feasible, effective and stable.

Our future work includes: study on problem space, improvement of performance, discovery of *h*-DESAR on data streams, and application of web mining or firewall log mining.

References

[1] Agrawal R, Imiclinski T, Swami A. Database mining: A performance perspective [J]. IEEE Trans Knowledge and Data Enginnering, 1993,5: 914-925.

[2] Agrawal R, Srikant R. Fast algorithm for mining association rules [A]. Proceeding 1994 International conference Very Large Data Bases (VLDB'94).

[3] Jiawei Han, Micheline Kambr. Data Mining-Concepts and Techniques [M]. Beijing: Higher Education Press, 2001

[4] Y Fu and J Han. Meta-rule-guided mining of association rules in relational databases[C]. KDOOD'95, 39-46, Singapore, Dec 1995

[5] Jie Zuo, Changjie Tang, Zhang Tianqing. Mining Predicate Association Rule by Gene Expression Programming[C]. WAIM, 2002

[6] C. Ferreira. Gene Expression Programming: A New Adaptive Algorithm for Solving Problems[J]. Complex Systems, 2001, 13(2): 87~129

[7] Jie Zuo. Research on the Key Techniques of Gene Expression Programming: [Ph. D. dissertation]. Sichuan: Sichuan University, 2004

[8] Silberschatz, Korth. Databse System Concepts, Fourth Edition, McGraw-Hill Computer Science Series, 2001

[9] DE CASTRO L N, VON ZUBEN F J .Artificial Immune Systems: Part I-Basic Theory and Applications[J].Technical Report, TR- DCA Ol/99, 1999, 12.

[10] Dasgupta D., Ji, Z., Gonzalez, F.. Artificial immune system (AIS) research in the last five years [J]. Evolutionary Computation, 2003. CEC '03.

[11] S. Forrest, A. S. Perelson. et al. Self-Nonself Discrimination in a Computer. In Proceedings of IEEE Svmposiimi on Research in Secwitv and Privacy, 1994.

[12] Tao Li, Xiaojie Liu, and Hongbin Li. A New Model for Dynamic Intrusion Detection [C]. CANS 2005, LNCS 3810, pp. 72–84, 2005.