Research on Evolutionary Immune Mechanism in KDD

Yiqing Qin[1, 2] Bingru Yang[1] Guangmei Xu[1] Wei Hou[1]

[1]College of Information Engineering, University of Science and Technology, Beijing 100083, P .R .China

[2]Department of Computer and Automation, Beijing Institute of Machinery, Beijing 100085, P.R.China

Abstract

In this paper, we introduce an evolutionary immune method to solve the problems from dynamic data mining. We first present a new concept to describe our specific problem domain. Then, we identify the evolutionary immune mechanism in KDD by illustrating how the elements involved in the domain can be modeled as the ones in an immune model. Accordingly, we describe a dynamic mining algorithm in detail and prove it correct and effective. Finally, we conclude our work and present proposals for future work.

Keywords: Dynamic data mining, Immune algorithm, Evolutionary immune mechanism, Pattern mining

1. Introduction

Many KDD systems need to explore precise information from regularly changed data. In such systems, we have to maintain the knowledge dynamically because either frequent or occasional data updating may change the rules explored before.

Some research on association rules' updating have proposed several algorithms. On the basis of the D.W.Cheung's research on rule maintenance and updating [1], He presented SEA algorithm [2];Feng proposed the IUA/PIUA algorithm [3]; Song, Yang and Ji gave a series of algorithms for the updating [4]-[5]. Most of the above algorithms for dynamic data mining are called as incremental mining algorithms because they update the rules based on data increment. They usually depend on the static mining methods such as Apriori [6] to update and maintain association rules, obtaining more precise mining results.

However, mining at a new time point may inevitably produce occasional or random data so as to have an impact on the results. In addition, with the data increasing, the structure of knowledge base being usually stable has to change. Thus , it is necessary to survey the knowledge base as a whole rather than to choose the rules only according to one mining process.

We introduce an evolutionary immune method to solve the above problems since an immune system has remarkable information-processing abilities. It learns to recognize relevant patterns, remember patterns that have been seen previously, and use combinatorics to construct pattern detectors efficiently. Also, the overall behavior of the system is an emergent property of many local interactions [7].

The remaining of the paper is organized as follows. In Section 2, we give a new concept to describe the specific problem domain; in Section 3, we propose the evolutionary immune mechanism in KDD by comparison between biological immune process and dynamic mining process; we accordingly present an evolutionary immune algorithm for dynamic data mining in Section 4 and prove it effective and correct by experiment in Section 5; we conclude our work in Section 6.

2. The dynamic mining process based on data increment

The dynamic data mining is referred to as a mining with motion, which dynamically examines the patterns each mining process and summarize them to evaluate the knowledge based on static mining. Actually, the dynamic data mining comprehensively evaluates the knowledge to decide what to be discarded from the view of the knowledge evolution. It requires KDD systems to take into account not only the dynamic database but also the dynamic knowledge base when making incremental mining.

Definition 1. The environment space of dynamic data mining is a tuple with five elements $\Omega = (U, V, R, S, T)$，where U stands for a dynamic database，V for a dynamic knowledge base,，R for the domain of interestingness；S for varying time and space；T for the set of domain and threshold.

For dynamic mining, a database with a time attribute T will be divided into several logical parts each of which is noted as DB_i according to the time interval Ti which is defined as the following: T_i = {tractime$_i$, tractime$_j$}where tractime$_i$ < tractime$_j$ if $i<j$ and $i, j = 1, 2, \ldots, n$. If the database has no time attribute, it will be logically divided into n parts each

of which can be represented as DB_i, satisfying $DB_i<DB_j$, if $i<j$ and $DB_n = DB$.

By the database division , the dynamic mining process is presented and described as the following:

Definition 2. The dynamic data mining based on data increment is referred to as a dynamic mining process if it keeps on mining the logical parts of the database in accordance with time sequence on the basis of logic division of the database and synthesizes the mining results to evaluate the knowledge mined.

The dynamic mining process uses not only incremental mining methods to finish a mining procedure but also each mining result to carry on dynamic analysis, preventing the data from the impacts of undetermined factors and conducting each mining. However, the dynamic mining process is different from the incremental mining on that it emphasizes on comprehensive and historical evaluation on mining results rather than concrete research on algorithms. The goal of the former is to use not only the last mining results just before current mining but also each mining ones to carry on rule analysis, while the latter is to utilize the last results to lower current algorithm's complexity to obtain effectiveness and efficiency. The dynamic mining process pays attention to following closely the development of the knowledge or the rules to explore the interested ones in stead of only mining the concrete results, avoiding interference on the results caused by causal factors possibly existing in one mining process.

The dynamic mining process is a macro view on the basis of each mining which is considered as a micro process. This characteristic makes it possible to solve the problem of the dynamic mining by the evolutionary immune mechanism.

3. The evolutionary immune mechanism in KDD

3.1. Biological immune process and AIS

In a biological system, an immune response is caused by an antigen invading. The immune molecules recognize antigens and pass the information to the active immune cells; then antibodies are selected and mutate to proliferate and differentiate according to the affinity between the antibodies and the corresponding antigens; finally, the humoral immunity is formed . The biological system can also remember what has happened and quickly respond to the same antigens next time.

The characteristics of the biological immune system(BIS) suggest people to deal with the engineering applications in the same way. The Artificial Immune Systems (AIS) are created to solve practical problems. Immune Network [8]-[9], Clonal Selection [10]-[11] and Negative Selection [12] are three of the most popular AIS models. Although many kinds of AIS are created from different aspects of BIS, they commonly build models for the affinity between antigens and antibodies.

3.2. The evolutionary immune mechanism in KDD

Figure 1 illustrates the similarity between the biological immune process and the dynamic mining process by the comparison between them.

- "antigens" corresponds to the new coming data in dynamic data mining;
- "antibodies" corresponds to the knowledge mined;
- "memory library" corresponds to the knowledge base;
- "producing the immune molecules" means forming the initial population from memory library or forming it randomly;
- "antigens meet antibodies" means the evolution of the knowledge after new data coming;
- "antibodies identify antigens" means to evaluate the adaptability of antibodies by the degree of correlation to the knowledge or the rule;
- "population of the antibodies" means that antibodies will be selected and mutate to obtain new population based on the last mining;
- "immune memory" means to reserve the antibodies as rules which are matching to the antigens, ready for knowledge presentation or the generation of the initial population.

We have presented KDTICM(Inner Cognition Mechanism of Knowledge Discovery Theory) to solve the problems of data mining[13].In terms of KDTICM and the similarity between the biological evolutionary immune process and the dynamic mining process on knowledge discovery, the evolutionary immune mechanism in KDD is proposed with the following characteristics:

- New data and old knowledge are considered as antigens and antibodies respectively. Antibodies proliferate, mutate and differentiate in accordance with the affinity between the antigens and antibodies;
- By the heuristic coordinator presented in KDTICM, antibodies inoculate in which the commonsense and the knowledge of the users and experts are referred to as vaccine. That means the adaptability of the antibodies are

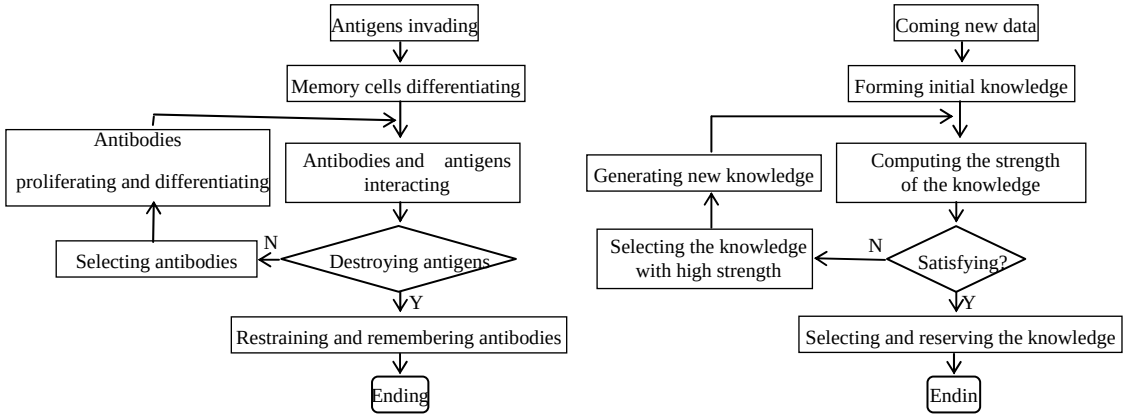

Fig. 1: The similarity between biological immune process and dynamic mining process.

improved to promote the ability to obtain new knowledge through directional mining.

- The new patterns obtained are evaluated first to expect to discard the contradiction except the repeated or redundant ones rather than reserved directly to the memory libraries. The repetitions are more possible to be explored when exploring the initial population to form a new response. It is easier to maintain the knowledge bases in real time by that way.

To dynamic mining process, it is significant to do research on knowledge discovery based on the evolutionary immune mechanism. The dynamic mining process aims at quick and effective mining, by identifying new data and utilizing the historical knowledge and their parameters. By immune memory theory, the knowledge repeatedly appearing are protected and kept rather than discarded only because it does not meet the threshold in some mining process before. This will prevent the patterns obtained from the impacts caused by data's random distribution .

4. Algorithm

A data mining on time series database can be regarded as a dynamic mining process. It not only uses an incremental mining algorithms to get the patterns each time, but also comprehend and analyze dynamically the patterns to discard the ones which is influenced by undetermined factors, improving and conducting the mining process next time. This is an evolutionary procedure.

4.1. Definitions for knowledge representation and affinity measure

4.1.1. Persistency of a rule

Definition 3. For a rule R, let $0<\lambda<1$， the following

$$\alpha = q_{i,n}+\lambda(q_{i,n-1}+\lambda(q_{i,n-2}+\ldots+\lambda(q_{i,2}+\lambda q_{i,1}))) \quad (1)$$

is persistency of R relative to the ith parameter.

The persistency of a rule shows the persistent degree of it relative to the ith parameter. In practice, λ can be set as required. Since $0<\lambda<1$， λ^n will tend to be zero with the mining times increasing, showing that the impact of the patterns mined before on the current mining decreases.

Definition 4. In the dynamic mining process ,if the current mining is the nth one, then the persistency of rules mined by $n-1$ times before is

$$\beta_i = \alpha_i/((1-\lambda^n)/(1-\lambda)) \quad (2)$$

β_i shows the persistent degree of a rule till now. The threshold of the rule's persistency can be predefined to determine whether a rule be kept or discarded.

4.1.2. Knowledge representation

Knowledge representation is necessary for evolutionary immune system. Here, antigens and antibodies are represented in real numbers.

Assume s is the number of the attributes in a transaction database, X_i is the i^{th} attribute, D_i is the domain of X_i,where i = 1, 2, …, s, then a record of the database corresponds to a vector with s dimension, its components on each dimension corresponds to the range of the attribute. One of its orthogonal division of subrange is $D_{i,1}$, $D_{i,2}$, …, $D_{i,t}$, then a mapping can be defined as $D_{i,j}\rightarrow j$, j = 1, 2, …, t_i.

A record can be represented by the mapping mentioned above: if $D_{i,j}$ is the i^{th} attribute of the record, then the i^{th} bit of the codes of the record can be set to j; if equal to zero, then to zero. Thus , the i^{th} bit of the code of a record is a determined value among 0, 1, …, t_i, namely, the range of the i^{th} bit of a record code is {0, 1, …, t_i}, noted asV_i, namely V_i = {0, 1, …, t_i}.

Definition 5. The Cartesian production of the range

of the codes V_i corresponding to all attributes in a transaction database

$$V = V_1 \times V_2 \times \ldots \times V_s \quad (3)$$

is the form space of the antigens.

The representation of the antibodies(the frequent itemsets) is the same way as the antigens'.

4.1.3. Similarity relationship between patterns

Definition 6. The patterns corresponding to the two time series $\boldsymbol{A}_1 = \{x_{1,1}, x_{1,2}, \ldots, x_{1,n}\}$ and $\boldsymbol{A}_2 = \{x_{2,1}, x_{2,2}, \ldots, x_{2,n}\}$ are $\{a_{1,1}, a_{1,2}, \ldots, a_{1,n-1}\}$ and $\{a_{2,1}, a_{2,2}, \ldots, a_{2,n-1}\}$ respectively. If $d(\boldsymbol{A}_1, \boldsymbol{A}_2) \leqslant \mathrm{pc}$ is satisfied, then $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are two similar temporal serial patterns within a prescribed error pc.

The similarity relation described above does not satisfy transitivity which is usually considered as a necessary characteristic for a similarity relation. For coding, the definition is improved to satisfy the similarity relation with transivity.

Definition 7. The interval between –90° and 90° is divided into n equal parts. If the code of a line is i, then it is similar to the pattern corresponding to the angle $-90° + 180°/2n + i \times 180°/n$. If a line's code is equal to the other's, then the two lines are similar.

4.1.4. Affinity measure

Traditional methods used to compute the affinity between antibodies and antigens include hamming distance in hamming space, Euclidean distance in Euclidean space and Manhattan distance in Manhattan form space. According to the defined representation for antibodies and antigens and similarity relationship between patterns, the affinity between antibodies and antigens is defined as the following:

Definition 8. $\mathrm{Ag} = (\mathrm{ag}_1, \mathrm{ag}_2, \ldots, \mathrm{ag}_L)$ is the code of an antigen and $\mathrm{Ab} = (\mathrm{ab}_1, \mathrm{ab}_2, \ldots, \mathrm{ab}_L)$ is the one of an antibody, let

$$\delta = \begin{cases} 1 & \text{if } \mathrm{ag}_i = \mathrm{ab}_i, \\ 0 & \text{else,} \end{cases}$$

where $i = 1, 2, \ldots, L$,then the affinity of Ab to Ag is

$$W = \sum_{i=1}^{L} \delta_i \Big/ L \quad (4)$$

That the value of W is equal to one means the complete matching between the antibody and antigen. It can imply the support to the series when computing the affinity.

Definition 9. The sum of the affinity of an antibody to all antigens is called as the stimulation degree of all antigens to the antibody.

The stimulation degree is critical to the decision on accepting or rejecting an antibody.

4.2. Algorithm

The dynamic data mining algorithm based on evolutionary immune mechanism is described as the following:

Input: n(dividing the interval between –90° and 90° into n equal parts);window w;time-series $\boldsymbol{A}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\ldots, \boldsymbol{x}_n\}$.

Output: the temporal sequential pattern

Step1: obtaining the corresponding serial set $\boldsymbol{W}(s)=\{s_1, s_2, \ldots, s_{n-w+1}\}$ according to the time series and the value of the window w;

Step 2: Computing patterns of each serial in the set as the antigens Ag and encoding them;

Step 3: Selecting randomly some antigens from Ag to form an initial antibody population Abs and put them into memory matrix $\boldsymbol{M}$;

Step 4: Computing affinity: computing the affinity of each antibody Ab in $\boldsymbol{M}$ to the current antigen Ag and getting the support;

Step 5: Clone selection: cloning the antibodies with high affinity;

Step 6: Antibody maturity: The antibodies mutate and inoculate vaccine;

Step 7: Reselecting: computing the affinity of each Ab to the current Ag, reselecting the antibodies with the higher support and regarding the ones with the lower support as the new current antigen;

Step 8: Executing from step 5 to 7 repeatedly till the prescribed running times is satisfied;

Step 9: If computing the first time, then going to step 11;

Step 10: Inhibiting the antibodies: inspecting the new coming antibodies whose support exceeds the minimum;

Step 11: Classifying and remembering the antibodies with high support and their persistency;

Step 12: Exhibiting the patterns whose support is higher than the minimum;

Step13: Keeping watching on new data coming;

Step 14: Generating a new serial set $\boldsymbol{W}(s)=\{s_1, s_2, \ldots, s_{n-w+1}\}$with new data;

Step 15: Forming a new initial antibody population Abs from the memory set and putting them into the matrix $\boldsymbol{M}$, going to Stpe4;

Step 16: Repeating from step 4 to step15 till the condition is satisfied;;

Step 17: Ending.

Experiment

Experiments are done on the time series databases to prove the effectiveness of the evolutionary immune mechanism of knowledge discovery in practice.

In the experiment, the average prices of a certain stock during eight months are used as the data set to explore sequential patterns in continuous four days, namely w is set to four. And set n to 9, the maximum of error to 20°, the support to 0.15.

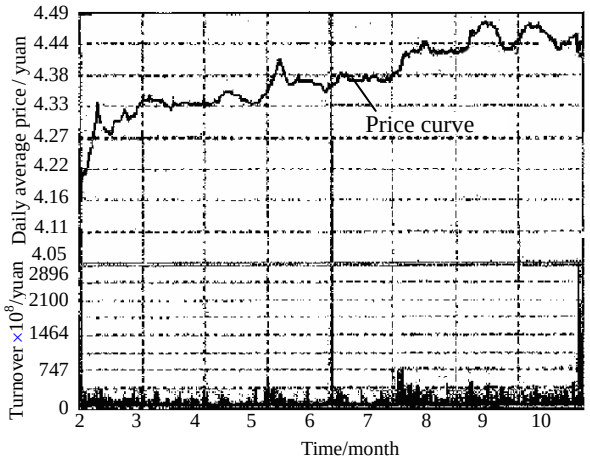

Fig. 2: The average price trend for a certain stock.

Two sequential patterns are obtained when experimenting on data of first three months(Fig.3 (a) and (b)). The patterns show a rising and modulating state of the stock.

Next, the experiment on the data of two more months quickly shows that the first two patterns are still frequent ones.

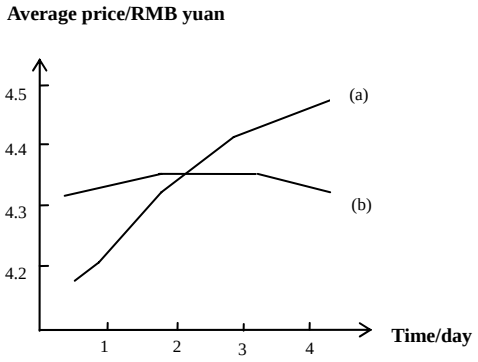

Fig. 3: Patterns shown at the first and second runtime.

Thirdly, the experimental data of other three months are added and the result shows another declining pattern(Fig.4 (c)) besides the first two frequent patterns. Since the new one appears late, the trend of the stock is still rising without regard to the third declining one.

The dynamic mining process based on incremental data utilizes the evolutionary immune mechanism to mine and evaluate the patterns. It mines the patterns more quickly by the found patterns and their variations. It also evaluates the patterns by each mining rather than only by the last one. The evaluation from historical and comprehensive view prevents mining from disturbance caused by random factors.

5. Conclusions

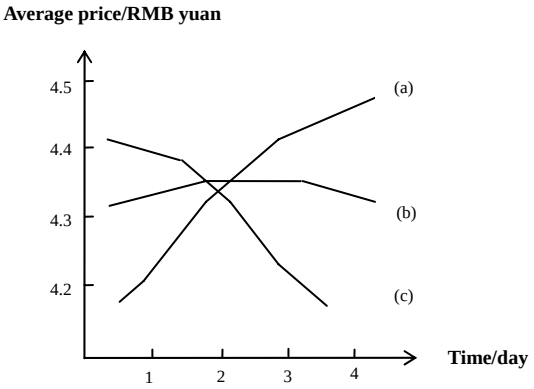

Fig. 4: Patterns shown at the third runtime.

According to the characteristics of dynamic data mining, we propose the evolutionary immune mechanism in KDD on the basis of AIS models. We have defined several significant concepts like knowledge representation and affinity measure. We also have described a dynamic mining algorithm in detail and proven it correct and effective. In the future, we will strengthen the theoretical basis of the evolutionary immune mechanism in KDD. We will also explore more possible applications for our approach.

Acknowledgment

This work is partially supported by National Natural Science Foundation of China (Grant No. 60675030).

References

[1] D.W. Cheung and J. Han, A fast algorithm for mining association rules. *Proceedings of the 4th International Conference on Parallel and Distributed Information System*, pp. 73-84, 1996.

[2] Y.X. He,G. Zhang and L. Shi, Maintenance on Association Rules. *Computer Engineering and Application*, 10: 203-205, 2002.

[3] Y.C. Feng and J.L. Feng, Incremental Updating for Association Rules. *Transactions on Software*, 4: 301-306, 1998.

[4] Y.Q. Song, Y.Q. Zhu and Z.H. Sun, Mining and Updating the Largest Frequent Itemset Based on FP-tree. *Transactions on Software*, 9: 1586-1592, 2003.

[5] G.L. Ji and M.Yang, Quick Updating for the Largest Frequent Itemset. *Transactions on Computer*, 1: 128-135, 2005

[6] R. Agrawal and Srikant, Fast algorithms for mining association rules. *Proceedings of the*

20th International Conference on Very large Databases, pp. 487-499, 1994.

[7] D. Dipankar, Advances in Artificial Immune Systems. *IEEE Computational Intelligence Magazine*, 11: 40-49, 2006.

[8] E. Hart and P. Ross, Exploiting the analogy between immunology and sparse distributed memories: A system for clustering non-stationary data. *Proceedings of the 1st International Conference on Artificial Immune Systems(ICARIS)*, pp. 49-58, 2002.

[9] J. Timmis, M. Neal and J. Hunt, Data Analysis with artificial immune systems and cluster analysis and kohonen networks: Some comparisons. *Proceedings of IEEE International Conference on Systems and Man and Cybernetics*, pp. 922-927, 1999.

[10] L.N.de Castro and F.J.V. Zuben, aiNet: An artificial immune network for data analysis. *Data Mining: A Heuristic Approach*, Idea Group Publishing, USA, pp. 231-259, 2001.

[11] De Castro and V. Zuben., An evolutionary immune network for data clustering. *Proceedings of the IEEE Computer Society Press SBRN001*, pp. 84-89, 2000.

[12] http://www.cs.unm.edu/~immsec/publications/virus.pdf.

[13] B.R. Yang, *Knowledge Discovery and Knowledge Engineering*, Beijing: Metallurgical Industry Publishing House, 2000.