Emergencies prediction model based on efficient mining for micro-blog content

Wenxia Bi

School of Art and Design, Jiangxi Science and Technology Normal University, Nanchang, 330013, China

wenxiabi@yeah.net

Keywords: association rules; emergencies; gene expression programming

Abstract. In the case of the increasing amount of micro-blog information data, traditional mining techniques used to forecast emergency have tedious defects, like time-consuming and poor stability and so on. To solve these problems, an emergencies predictive method of efficient algorithm for micro-blog content mining is proposed, combining niche technology and gene expression programming in the data mining technology, through penalty function to set support threshold, and search for stronger association rules during data mining, which benefits effectively avoiding premature algorithm, so as to solve the problem of redundant rules, realize emergencies prediction . Experiment simulation proved emergencies prediction method for efficient mining based on efficient micro-blog content has high accuracy and strong robustness.

Introduction

In the era of evolving information technology, the emergence of information technology will produce large amounts of data in different formats stored in the database. How to search emergencies information from a large amount of data is currently the focus problem pending to be dealt with [1-2]. Under normal circumstances, artificial data mining method for emergency prediction is cumbersome, and have strong constraints, thus, resulting in multiple problems like time-consuming and low precision [3-4]. To compensate for these shortcomings, an efficient mining technique for micro-blog content is utilized to build emergencies predictive methods, combining niche technology and gene expression programming in the data mining technology, through penalty function to set support threshold, and search for stronger association rules during data mining, which benefits effectively avoiding premature algorithm, so as to solve the problem of redundant rules [5-6]. The simulation proved emergencies prediction method for efficient mining based on efficient micro-blog content has high accuracy, and meet the requirements of emergencies forecasting.

Emergencies prediction principle of efficient mining based on micro-blog content

Mining for information related to emergencies in micro-blog content

The micro-blog content efficient mining techniques are defined as the technology through related data to mine information, extracting valuable information from a large, fuzzy micro-blog content. Association rules GEP algorithm Data in data mining techniques includes self-adaptive evolutionary algorithm of genotype and phenotype, which is characterized by simple, convenient, stable and fast. Because GEP algorithm is random search algorithm derived from natural selection and genetic mechanisms of the nature, so there will be difficult to obtain global optimal solution. If niche technology in data mining techniques is used to supervise and predict emergencies by GEP algorithm, under the premise of ensuring the convergence rate, but also enhance the accuracy of the optimal solution, eliminating the unbalance problem produced by traditional GEP algorithm of mining techniques in the process of changing, and population diversity of GEP algorithm is preserved in the process of change. Specific steps are described as follows:

(1) First of all, data analysis is performed for emergencies forecast information, SAS statistical software is applied to identify statistical significance of large area emergencies prediction.

(2) Large area emergencies prediction information database is determined; randomly generated initial population is allocated to each niche evenly, perform the evolution and fusion algorithm within niche territory, so as to complete genetic manipulation of GEP algorithm, like crossover and mutation, under the premise of in line with emergencies prediction constraint requirements to output strong association rules.

(3) The design of the fitness function of emergency prediction will interfere with the performance of the GEP algorithm, therefore support and confidence of emergencies prediction is important criterion to measure strength of association rules of micro-blog content in efficient mining process.

Dynamic fitness function $f(j)$ of emergency prediction is defined as follows:

$$f(j)\begin{cases} spt(j)-\partial(l)\times\min\_spt+\begin{pmatrix} cnf(j)-\min\_cnf \\ cnf(j)-\min\_cnf\rangle 0 \\ spt(j)-\partial(l)\times\min\_sp\rangle 0 \end{pmatrix} \\ 0 \\ spt(j)-\partial(l)\times\min\_spt\geq 0 \\ cnf(j)-\min\_cnf\geq 0 \end{cases} \tag{1}$$

Where, $\partial(l)$ is the penalty coefficient, when mining association rules, the degree of interest for item sets varies along the length of item sets, so the minimum support threshold is set as penalty coefficient $\partial(l)=\frac{l}{2^{l-1}}$. Among them, l is the length of item sets, $spt(j), cnf(j)$ are support and confidence of j-th rule formed through genetic manipulation; $\min\_spt, \min\_cnf$ are the pre-set minimum support threshold and minimum confidence threshold.

During the process of micro-blog content efficient mining, the selection of recombination probability would interfere with convergence of GEP algorithm. The greater the recombination probability, the faster the changes of new combined group is, the damage of genetic scale is deeper, but if the recombination probability is too small, the search process will be slow, so the algorithm evolutionary is stagnating. So, during efficient mining process of micro-blog content, the average fitness is mapping the direction of movement of the overall population, according to the average fitness, recombination probability pc is defined as follows:

$$pc=\begin{cases} pc*\dfrac{f\max-f(j)}{fmanx-favg} & f(j)\geq fmanx-favg \\ pc & f\langle favg \end{cases} \tag{2}$$

During efficient mining process of micro-blog content, $f(j)$ is the individual with larger fitness function value between two individuals participating in the operation, $f\max$ is the maximum fitness value, $fmanx$ is the average fitness value.

In summary, the basic principles of efficient mining micro-blog content can be obtained, because it converges fast, and has strong stability, the universal application is strong.

Realization of emergency prediction

In order to improve the prediction accuracy, the principal component analysis and micro-blog content efficient mining methods are combined to form a prediction method of combination \ data mining, the specific expression formula as follows:

In emergencies prediction method based on efficient mining micro-blog content, the impact factor is $\{x_1, x_2, ..., x_i ....., x_m\}$,

x_i represents the i-th factor, y represents the prediction level, the representation of predicted level data is $[y_i]=[x_{i1}x_{i2}....x_{im}]$, mathematical formula predicted at the time period i is $\hat{y}_i = f(x_{i1}, x_{il}, ..., x_{im})$.

During the process of emergencies prediction based on efficient mining micro-blog content, the steps for emergencies forecasting is:

To obtain the emergency prediction formula. Assuming predicted emergencies data of n periods, the number of impact factor is m , the prediction formula is:

$$x'_{ij} = \frac{x_{ij} - x_j}{\sqrt{\mathrm{var}(x_j)}}, (i = 1, 2, ..., n, j = 1, 2, ..., m) \quad (3)$$

Thereinto,

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad (4)$$

$$\mathrm{var}(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2 \quad (5)$$

covariance matrix R of normalized data is calculated

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \vdots & \vdots \\ r_{m1} & \cdots & r_{mp} \end{pmatrix} \quad (6)$$

（3）In emergencies prediction process based on efficient mining micro-blog content, the eigenvalues and eigenvectors of related matrix are calculated, and sequenced according to the size of eigenvalues, that is $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$

（4）In emergencies prediction process based on efficient mining micro-blog content, the contribution rate of each main component is calculated, the cumulative contribution rate of first p principal components is $\sum_{i=1}^{p}\lambda_i / \sum_{i=1}^{n}\lambda_i$. Under normal circumstances, the cumulative contribution rate of the first p principal components is 85%, which can be utilized to determine the information obtained when the first p principal components factor replace the initial impact factor. In the case of $p \langle m$, the principal component analysis method can simplify the input information of model, eliminate redundant information, so as to reduce dimensionality.

In summary, emergencies prediction results based on efficient mining micro-blog content can be obtained, according to the prediction results can complete emergencies forecasting.

Experimental simulation analysis

In order to verify the effectiveness of the proposed emergencies prediction model based on efficient mining micro-blog content, it needs an experiment. More than 2500 microblogs of Tencent microblog are as the initial data to establish the simulation environment in the MATLAB environment. A virtual emergencies prior model was established, using artificial data mining algorithm and efficient mining algorithm for microblog content to conduct an experiment respectively. In the course of the experiment, 10 experimental results are selected to analyze emergencies. The ratio of the emergencies prediction results obtained from the mining data based on microblog content with the occurrence of actual emergencies is as the reference point for emergencies prediction method, which is called the accuracy rate of prediction, the formula is as follows:

$$F = \frac{f_a}{f_{all}}$$

Where, f_a is the results of emergencies prediction based on microblog content to mine data, and f_{all} is the occurrence of actual emergencies.

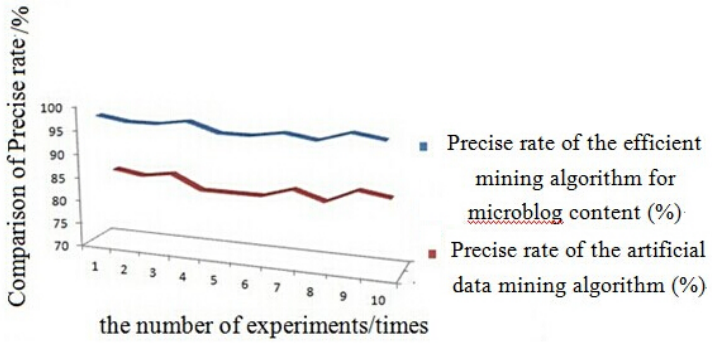

Figure 1 Trend graph of precise rate comparison with different algorithms

According to the figure above to obtain the following table:

Table 1 Data tables of precise rate comparison with different algorithms

The number of experiments (times)	Precise rate of the efficient mining algorithm for microblog content (%)	Precise rate of the artificial data mining algorithm (%)
1	98	85
2	97	84
3	97	85
4	98	82
5	96	82
6	96	82
7	97	84
8	96	82
9	98	85
10	97	84

According to the experiment above, it can be learned that the proposed method used for emergencies prediction can greatly improve the accuracy of forecasting, and has strong robustness.

Conclusion

For the defects like long time-consuming and poor stability produced by the traditional artificial data mining algorithm in the process of establishment of emergencies prediction model, emergencies prediction method for efficient mining based on efficient micro-blog content is proposed, according to a mining association rule algorithms which combining niche technology and gene expression programming to perform emergencies micro-blog information mining. Support threshold is set by the penalty function. According to mined micro-blog content which related to emergencies to get emergencies prediction results. The experimental results show that the proposed algorithm for mining emergencies related micro-blog content, and get emergencies prediction results by the mined information, can greatly improve accuracy of emergencies forecasting.

References

[1] Tan Hengui, Wang Wenjie, Li Youhua. Review of Classification Algorithm in Data Mining［J］. Microcomputer & Its Applications, 2005（2）：4-6.

[2] Huang Xiaofang. Algorithm of Decision Tree in Data Mining and Its Application［J］. ORDNANCE INDUSTRY AUTOMATION 2005, 24(2):35-36.

[3] Luo Ke, Lin Mugang, Xi Dongmei. A Review on Classification Algorithms in Data Mining［J］. Computer engineering, 2005（1）: 8-10,16

[4] Liu Zhao, et al. The Research of Data Mining Based on Neural Networks［J］. COMPUTER ENGINEERING AND APPLICATIONS, 2004,（3）：172~173.

[5] Zheng Zhuoyuan, Zhou Ya. Influence of Data Mining on Information Security [J]. MODERN COMPUTER, 2008(03):36-39.

[6] Luo Yueguo. A Design of Intrusion Detection Model Based on Data Mining [J]. Journal of Xi'an university of arts and science (natural science edition), 2010(03):112-113.