Crisis prediction in e-learning through data mining technology: an empirical investigation

Ke Zhu, Jin Zhang[*]

Department of Information Technology, Henan Normal University, East of Construction Road, Xinxiang, China

Received 1 March 2014, www.cmnt.lv

Abstract

Crisis warning is a kind of semi-structured or unstructured problems with a lot of uncertainties. In order to examine learners’ actual achievements and timely warning the problems of the students' learning, crisis prediction techniques are imperative as they assist the teachers in monitoring learners’ progress and, determining their development and competencies. Many factors have no historical data and corresponding statistics, therefore crisis prediction is difficult to calculate and evaluate scientifically. There are many conventional methods of analysis has a lot of limitations and the results are not accurate enough. In this paper, a crisis prediction technology method for e-learning courses, based on data mining techniques and detailed student data, is proposed. An empirical field experiment involving 129 university students was conducted. The results were found to be significantly better than those reported in relevant literature.

Keywords: crisis prediction, data mining, big data analytics

1 Introduction

Information and Communications Technologies have been considerable advances in the last few years. Technological promises to change learning have been around for years. As a new form of learning style, Electronic Learning (e-learning) in virtual learning environments has grown in recent year [1]. The nature of e-learning can make the learner learning at any time and any place, also interact with numerous peers in internet. E-learning also benefits educational institutions have the opportunity to provide more students with course, reducing the times and expenditure. With learners increasingly mobile and working in remote locations, individuals will rely far more on e-learning [2].

Despite the numerous advantages, increasing the effectiveness and success rates of e-learning is progressing relatively slowly in many cases, due to several concerns und barriers on technology, pedagogical and learning environment levels. The key challenges of e-learning are also surfaced: difficulty in evaluating students’ work and intrigued with the potential of technology to help improve student learning.

In order to examine learners’ actual achievements and timely warning the problems of the students' learning, crisis prediction techniques are imperative as they assist the teachers in monitoring learners’ progress and, determining their development and competencies.

Conventional crisis prediction technology is simpler to deploy, analyse, apply, and have lower costs, and therefore, make the overall certification process easier [3]. However the conventional crisis prediction technology usually employ a single set of data source to assess may lead to inaccurate assessment of students’ state. Data mining technology is a hot issue in the field of artificial intelligence and database research, it can automatically analyse massive data of warehouse data, mining a rich and objective knowledge of crisis prediction, and applied to the crisis warning. Applying data mining techniques in the field of crisis warning has important theoretical signifycance and application value [4].

We mainly undertook the following works:

1) Proposed crisis prediction mechanism based on data mining technology.
2) Designed a representation of the domain knowledge constraints-based crisis prediction knowledge, which is based on the analysis of association rule mining method.
3) Designed the methods of crisis prediction and iterative warning method to generate crisis prediction information based on excavated warning knowledge.

2 Literature review

Several studies focus on the field of crisis prediction as follows:

1) Logistic regression analysis (LRA).

Logistic regression analysis is a kind of statistical method, it has been widely applied to the fields of crisis prediction, the specific method is to descriptive statistics and simple indicators test by selecting sample and defining variables, and then analysed the correlation between

[*] ***Corresponding author’s*** e-mail: zj@htu.cn

variables based on the test results, rejected highly correlated variables, and then logistic regression on this basis, then choose the best probability of closing value (split point), and closing value is the critical point of warning, test the obtained forecasting methods and results, finally we can get the final reliable model.

2) Multivariate discriminant analysis model (MDAM).

Multivariate discriminant analysis was a kind of statistical analysis to discriminant the research object performed category, score model originally proposed by Altman, in the generally form of its discriminant function is $Z = a_1x_1 + a_2x_2 + ... + a_nx_n$, whereby: Z is discriminant (discriminant value), $x_1, x_2, ..., x_n$ is a reflection of the characteristics variables of the study object, $a_1, a_2, ..., a_n$ is the determining coefficient for each variable [5].

Discriminant analysis must be known the category of the object and a number of variable values that indicate the observed characteristics, discriminant analysis are to screen the variables that can provide more information and to establish discriminant function to determine the warning critical value and forecast [6].

3) Warning evaluation method based on artificial neural network (ANN).

ANN is a network on the basis of physiological research in the brain, with some of the basic functional components simulation of biological neurons (i.e. artificial neurons), according to a variety of different ways of linking organized. Its purpose is to simulate some of the mechanism and the mechanism of the brain, through continuous learning in advance, can achieve self-learning function. Artificial neural network model is mainly based on BP neural network model [7]. BP artificial neural network is based on a multi-front to back propagation neural network algorithm. Since the transfer function of neurons using BP neural network is usually Sigmoid type differentiable function, you can achieve any nonlinear mapping between input and output, and has the most widely used in pattern recognition, risk assessment, adaptive control, which is widely used in the evaluation of crisis prediction indicators currently. Historical data on crisis prediction indicators is relatively small, in the case of non-linear change; the artificial neural network method is available for crisis prediction indicators self-learning evaluation.

4) Intelligent Crisis prediction Support System (IEWSS).

Intelligent Crisis prediction Support System (IEWSS) is an important branch of the decision-making system, along with neural networks, case-based reasoning, fuzzy reasoning, rule-based reasoning technology gradually entering the field of crisis prediction, brought new theories and methods to knowledge representation and reasoning of intelligent crisis prediction system [8]. Among them, the case-based reasoning technology is more widely used in the field of crisis prediction. Case-based reasoning first describe warning object characterization, based on these characteristics, retrieve similar cases from the case base, relatively the similarities and differences between the new issue and the old cases. Be adjusted through crisis prediction information and case base stored information comparing achieve the purpose of crisis prediction.

5) Sentiment index (SI).

Sentiment index method is to use the time difference between the relevant economic variables between each other to indicate the movements of the economy, by constructing synthesis and proliferation index to achieve the purpose of monitoring the economic performance warning [9]. This method is divided into four steps: The first step is to determine the time difference between the reference benchmark one cycle, which is a critical step; the second step is to select indicators; third step is divided into first, sync, lagging indicators; first four step is first, sync, lagging indicators were compiled diffusion index and synthetic index. Division first and lagging indicators can be used to synchronize gray correlation method, fuzzy nearness Act and discriminant analysis method [10].

Diffusion index can be integrated volatility for each variable, can reflect macroeconomic volatility process, but also can effectively predict the turning point of the economic cycle, but the strength of the economic cycle change is not clear. Diffusion index value DI_t at time t is

$$DI_t = \sum_{i=1}^{I} w_i I\left[x_{it} > x_{i,t-1}\right] \times 100 . \tag{1}$$

6) Auto-Regressive and Moving Average Model (ARMA).

ARMA model is a time-series forecasting methods, proposed by Jenkins, and also called Box-Jenkins Model. ARMA model based on the basic idea is: except in very limited circumstances, between the observed value of chronological order arrangement, there have a dependent relationship or self-correlation among almost all of the time series, this self-correlation indicates that the continuation of the variable development, and this self-correlation once described quantitatively, it can predict future values from past values of the sequence [11].

The general form of ARMA model:

$$Y_i(t) = C_0 + \sum_{j=1}^{n} C_j AR(j) + \sum_{j=1}^{n} C_j MR(j) , \tag{2}$$

where: $Y_i(t)$ is t stage predictive value of the first indicator, n is the maximum lag phase, j is lag periods, k is moving average, C_j (j=0,1,2,3,…,k) is regression coefficients $AR_{(j)}$ is son regression model $MR_{(j)}$ is still moving average model. Use this method to predict the ARMA model has the following advantages, not only investigated the past values of the predictor variables, but the errors fitting generated by past values of model is also as an important factor enter into model; no need to pre-determined the development patterns of sequence, one can assume that the style may be applicable, the method itself will be in accordance with the prescribed procedures, approach to one of the best fitting equation by identifying modified, until you get a satisfactory model style; due to continue decomposition for the remaining items to make it meet the assumption of regression methods, so you can use mathematical statistical methods to confidence interval estimation for predict values. ARMA model applied to more widely range [12].

3 Materials and methods

In this research, the steps to build crisis prediction system are as follows:

1) Design for the representation of crisis prediction knowledge.

Warning knowledge representation method refers to the feasibility and validity of crisis prediction knowledge which represented by machines, it is the unity of a data structure and control structures, both consider the storage of knowledge, but also consider the use of knowledge. Traditional knowledge representation methods are: first-order predicate logic, production rules, semantic networks and frameworks.

2) Design for Warning knowledge base.

The knowledge representation, organization and storage mode of warning knowledge base will affect the efficiency of crisis prediction module, but also will affect the update and the rich of warning knowledge, and ultimately will affect the intelligence level of the entire warning system.

3) Historical data pre-processing.

Data mining have strict quality requirements for the data processing. Data pre-processing is critical in the process of data mining. According to statistics, in the process of a complete data mining, data pre-processing will spend about 70 percent time on it, while the back of the excavation work only take about 10% of the total workload. Data pre-processing mainly includes that data cleaning, integration, transformation and reduction. Data cleaning is to clean up data by filling vacancies values, smooth noisy data, identify, remove encourage points and solve the inconsistent; data integration is to merge multiple data sources into a consistent data storage; data conversion is to convert data into a form suitable for mining, such as attribute data is scaled so that a relatively small fall into a specific range; data protocol is data mining results without affecting the premise, by numerical aggregation, remove redundancy approach compressed data improve the quality of the digging mode, reduce the time complexity.

4) The design of warning knowledge discovery method.

Many intelligent crisis prediction knowledge discoveries can be expressed as the following six categories tasks: classification, estimation, prediction, affinity grouping or association rule, clustering, description and profiling. Among them, the first three are examples of directional data mining; the purpose is to find the value of a specific target variable. The task of non-directional mining is affinity grouping and clustering, in the case of a not defined specific target variable aimed at to reveal the structure of the data. Establish a profile may be directed, may also be non-directional data mining tasks. This paper has proposed a crisis prediction knowledge discovery algorithm which is based on Apriority mining algorithm framework.

5) Monitoring data reprocessing.

The process of monitoring data pre-processing is similar to the process of historical data pre-processing, as tasks required may differ in approach.

6) Design of crisis prediction methods.

According the design of crisis prediction knowledge, the necessary of problem solve to design the appropriate warning methods. For example, crisis prediction knowledge is production rules; it could take the forward chain reasoning, backward chain reasoning and bidirectional chain inference methods. We proposed a method of a step warning and iterative warning in accordance with the warning rules of design.

7) System implementation and integration.

The above method is implemented as a separate subsystem that provides the necessary interfaces for host system or other systems, integrated into the main control system. Including the provision of real-time control interface, integration of multiple incident response interfaces, to produce the desired message for collaborative process, achieve linkage warning functions.

Warning work is divided into two phases: the training phase and crisis prediction stage. In the training phase, the crisis prediction system accept the achievement training data of student, which automatically obtain the desired results of crisis prediction rules system, warning rules mining can adopt the rules which is warning rule representation and warning rule mining methods that proposed by this research. In the crisis prediction stage, the system release warning information according to the warning strategy designed, student achievement based on user input, the rules to be warning courses and training phase obtained, crisis prediction strategy can take one step crisis prediction method.

Data need to be converted to the desired form before the crisis prediction rule mining and pre-warning. Currently, the objects of association rules research mostly is transaction database, their attributes values limited to Boolean or enumeration type. The attributes of results database is mainly numeric attributes (percentile scores) and category attributes. To the end, the attributes of the relational database need to be converted. In this paper, the divide interval method will be divided into several categories, as the class attribute is converted to numeric attributes; the value attribute range is divided into several intervals. Following the below method to convert numerical attribute of relational databases: let a certain attribute of relational databases A_j has a region taking value of *k*, let be symbol *k* respectively corresponding to $A_{j1}, A_{j2}, ..., A_{jk}$. To convert the class attributes and values attribute of relational database unified into Boolean or enumeration type attributes. In general, this conversion requires the steps to experience missing achievement and multiple scores handling, subsystem conversion, data discrete, data integration and transformation.

4 Results and discussion

To achieve the above software environment of warning programs is as follows: operating system is Windows 7, using SQL Server2000 database management system. Record $D_{Training}$ as training data set, D_{Test} as testing data set, $R=\{r_1,r_2,...,r_n\}$ is the warning rule sets of algorithm Gen Crisis prediction Rules mining from $D_{Training}$. Defined warning accuracy of rule $r:X\to Y$ as $P(r)$, warning accuracy of rules set R as $PR(R)$.

$$P(r)=\frac{\left|\{T:X\cup Y\subseteq T,T\in D_{Test}\}\right|}{\left|T:X\subseteq T,T\in D_{Test}\right|}*100\% . \qquad (3)$$

This research was applied on three e-learning courses, each course's 129 students' score for the training data source, dug out of the rules number for the number indicators of evaluating mining results, crisis prediction accuracy of the rule set as quality indicators of the evaluation mining results, verify the effectiveness of the crisis predict-tion program.

Set the minimum support were 0.40, 0.35 and 0.30, the minimum confidence respectively were 0.85, 0.80, and 0.70, set warning item sets according to the course grade value; item order relationship set is determined by the order of the course study. Models and methods mentioned in this article on the reality of the data source can reach more than 63% of the warning accuracy rate, by a reasonable set of parameters, up to more than 87% of the warning accuracy. Experimental results show that the model and method are effective in practice.

In addition, we found that the dig out law is a statistical sense during the experiment, sometimes has a strong correlation on the learning content; this law can be applied to the actual data to achieve better warning effect. Analysis of the reasons, there may have a similar way in terms of thinking, learning and learning methods in these courses.

The experimental results on overall accuracy of the three kinds of courses are depicted in Figures 1, 2 and 3.

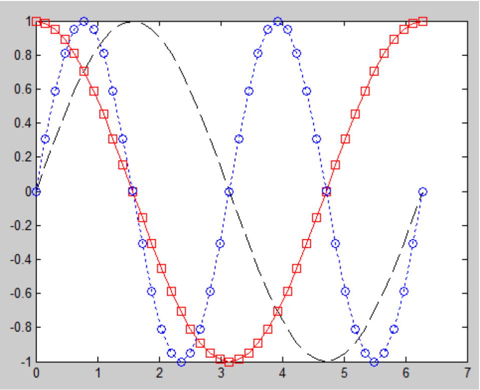

FIGURE 1 Overall accuracy results on the course 1

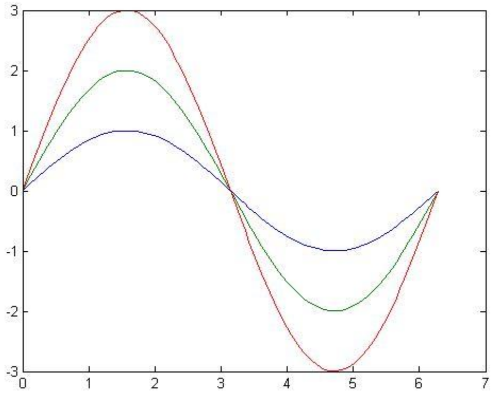

FIGURE 2 Overall accuracy results on the course 2

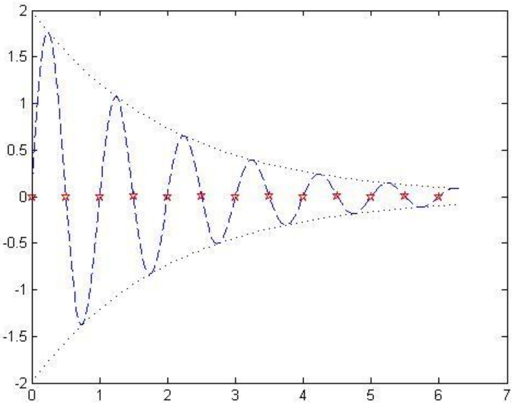

FIGURE 3 Overall accuracy results on the course 3

5 Conclusions

Based on preliminary empirical investigation, the following conclusions were obtained from this study.

A crisis prediction system for e-learning was proposed in the current study to reduce the risk of failure of learning, and the application of data mining was suggested as a method to manage the multi-variable endpoint data set.

Crisis prediction rules were effective in integrating the responses of the learner's data, minimizing experimental noise, and highlighting any distinct patterns within the real learning process.

After establishing the crisis prediction rules, discriminant analysis was employed to evaluate the capability of the library to detect and resolve different contaminants. Using the present library of responses, sets of earlier measurements could be classified with a high accuracy. These findings reflected the potential field application and capability of the crisis prediction system to assess learning quality in real time and rapidly detect any dropout.

Acknowledgments

This work is supported by soft-scientific project of Henan Province (No. 142400410425).

References

[1] Merigo J M, Casanovas M 2009 Induced aggregation operators in decision making with the Dempster-Shafer belief structure *International Journal of Intelligent Systems* **24**(8) 934-54

[2] Angeli C, Valanides N 2012 Examining the effects of text-only and text-and-visual instructional materials on the achievement of field-dependent and field-independent learners during problem-solving with modeling software *Educational Technology Research and Development* **52**(4) 23-36

[3] Chi M, Koedinger K, Gordon G, Jordan P, VanLehn K 2011 Instructional factors analysis: a cognitive model for multiple instruct-tional interventions *Proceedings of the 4th international conference on educational data mining*

[4] Fenton NE, Martain N, William M, Peter H, Lukrad R, Paul K 2007 Predicting software defects in varying development lifecycles using Bayesian nets *Information and Software Technology* **49**(1) 32-43

[5] Lattin J M, Carroll JD, Green P E 2003 Analyzing multivariate data *San Diego CA US Harcourt Brace Jovanovich* 123-6

[6] Getoor L, Mihalkova L 2011 Exploiting statistical and relational information on the web and in social media *Proceedings of the fourth ACM international conference on Web search and data mining ACM New York NY USA* 9-10

[7] Li S 2009 Research on learning styles and learning preferences Mining Model based on learning behavior *Wu Han Hua Zhong Normal University* (*in Chinese*)

[8] Yano T, Martins E, De Sousa F 2011 A multi-objective evolutionary algorithm to obtain test cases with variable lengths *Proceedings of the 13th annual conference on Genetic and evolutionary computation ACM* 1875-82

[9] Shermis M D, Hammer B 2012 Contrasting State-of-the-Art Automated Scoring of Essays *Analysis from* http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

[10] Guascha T, Alvarezb I, Espasaa A 2010 University teacher competencies in a virtual teaching/learning environment: analysis of a teacher training experience *Teaching and Teacher Education* **26**(2) 199-206

[11] Romero C, Ventura S, Pechenizkiy M, Baker R S J d 2011 Handbook of Educational Data Mining *CRC Press Taylor and Francis Group USA*

[12] Pardo A, Delgado K C 2011 SubCollaboration: large scale group management in collaborative learning *Software Practice and Experience* **41**(4) 339-465

Authors

Zhu Ke, born in 1982, Henan Province of China.

Current position, grades: lecturer in Henan Normal University
University studies: Master’s degree in educational technology in Henan Normal University in 2008.
Scientific interest: neural network algorithm, data mining.

Zhang Jin, born in 1983, Henan Province of China.

Current position, grades: lecturer in Henan Normal University.
University studies: Master’s degree in educational technology, Yunnan Normal University in 2008.
Scientific interest: learning analytics, data mining.