Research on college English student individual learning based on data mining model

Fengjuan Zhang

Zhengzhou university of industrial technology. Henan. China

Keywords: Data mining technology; college English student individual learning ;

Abstract. Data mining technology is generally to process, observe and analyze large-scale data by use of statistic analysis software, aiming to find the hidden rule or effective information existing in data. The application core of the data is large-scale college English student individual learning which completely depends on data source. With data mining technology as the entry point and in combination with the analysis on college English student individual learning characteristics, college English student individual learning implementation access based on data mining combination model has been explored in the thesis. Research on variable substitution to non-linear regression forecast model precision's influence, and seek the modelling method that can improve the forecast precision. Based on the Data mining, the transform in space and the weighted processing combined method, make full use of information that the primary data provide.

1 Introduction

With the advent of big data era, among the top 10 technologies that have been internationally recognized with the most powerful influence and development potential in the future, data mining technologies ranks the third as the further expansion and deepening of statistics and data base technology[1]. The rapid development of computer technology and the wide application of data base, people have seen a great progress in their data collection and storage ability; especially with the generalization of the internet in recent years, various data and information have begun to explode. In face of the mass and complex data, people tend to feel helpless and confused, hard to effectively analysis and deal with them, with some managers even making decisions purely by intuition, instead of making analysis and judgment based on the historical data, which is undoubtedly a loss[2]. It is under such circumstances that date mining technology comes out in response to the demands of the times.

2 Overview of data mining technology

Data mining refers to the process of mining the potentially useful and regularly existing information and knowledge that hides in the large magnitude of practical data that people are not clear about in the first place[3]. While judged from the perspective of finance, data mining is a whole new financial information processing technology, mainly characterized by the analysis and exploration for the mass and complex data in financial data base to discover the hidden key rule that could help individual or institute with investment decision, and thus help people to make accurate judgment or decision.

3 The characteristics of college English student individual learning

Big data requires a new process model to develop mass, high-growth-rate and diversified information assets with a stronger decision-making power, perception and procedure optimization ability[4]. college English student individual learning and analysis relies on data source, therefore, the characteristics of the data source also decides the characteristics of the big college English student individual learning .

3.1 Real samples rather than sampling

Cloud computation and data base can make it quite easy to acquire a sample data big enough and the entire data. Google can provide Google flue trend just because it has covered over 70% of American search market, no longer necessary to investigate the data by sampling but just to mine and analyze big data recording base. However the big data also have their flaws and the systematic deviation remains possible, for the real sample is not equal to the entire sample [5]. Therefore, there exists an issue on the threshold value of data scale. In case the number of data is less than the value, the questions can never be solves, while in case it reaches the value, there will come the solutions to the originally insurmountable questions[6]; even though the number exceeds the value, there won't be any more help to solve the questions.

3.2 Efficient rather than accurate

The traditional sampling requires a high accuracy in specific operation, because the slightest error may cause a grave consequence. Just imagine randomly selecting 1000 from the entire sample of the global 100mn people, in case of any errors in the operation on the 1000, there will produce a huge deviation among the 100mn[7]. While in case of full sample, the deviation will always be the same without the risk of magnifying. Google artificial intelligence expert Novarg ever wrote that the simple algorithm based on big data could be more effective than the complex algorithm based on small data. Data analysis is not simply for the sake of data analysis, instead it has many decision purposes and thus the timeliness also matters. Accurate calculation is conducted at the expense of time consumption, and in the era of small data, perusing accuracy is the forced method to avoid deviation expansion. In this big data era, rapidly acquiring a rough outline and development vein is far more important than the strict requirement for accuracy[8].

3.3 Relevancy rather than causality

Different from traditional logic reasoning, big data research requires a series of analysis & conclusion operations like statistic search, comparison, clustering and classification, and therefore, it inherits some characteristics from statistic science. Statistics pays attention to data relevancy or correlation. The so-called correlation means some rules existing between the values of two or more variables. "The analysis on correlation" aims to discover the correlation networking hidden in the data set which can generally be represented by support degree, reliability and degree of interest. The recommended algorithm of Amazon is quite well-known for telling what users might like by their consumption records which can be the historical records of the users or someone else. However, it cannot tell the reasons why they like. Just understanding the correlation is far from introducing the recommended algorithm to Amazon logistics and warehouse layout, or else some extra loss might be brought about. That is also the boundary line between relevancy and causality prediction.

4 College English student individual learning access based on data mining combination model

In health statistical research, it is necessary to discover a hidden rule from a lot of data, and it is best to present it in a mathematical model. Obviously the vast majority of these mathematical models are nonlinear. Because nonlinear regression models are more complex than linear regression models, it is not easy to calculate the regression parameters. On the premise of meeting the needs of the actual situation, sometimes non-linear models are approximated to regression models [9] to solve practical problems. By approximating a regression model by a nonlinear model, generally first substitute variables of the non-linear function and convert it into a linear model; afterwards, implement a linear regression, and then revert to the nonlinear model. Wherein, the calculation process of converting from a nonlinear model to a linear model, and then from a linear model to a nonlinear model, some interference information is added while the original information is lost, which will sometimes seriously affect the prediction accuracy of the nonlinear regression model obtained , whereas the combination forecasting model based on data mining methods can overcome this defect.

4.1 Principles and Methods

4.1.1 Form of the nonlinear mathematical model

The nonlinear mathematical model can be expressed as follows:

$$y = f(x_1, x_2, \dots x_m, a_1, a_2, \dots a_3 + e) \qquad (1)$$

where $e \sim N(0, \sigma^2)$. The independent variable x=(x1 ,x2, … , xm) $\in R^m$ in Model (1) is a point in a m-dimensional space; parameter $\alpha = (\alpha 1, \alpha 2, \ldots, \alpha l) \in Rl$ is a point in an l-dimensional space; the dependent variable $y \in Rl$ is a point in a one-dimensional space. Multivariate function $f(x1, x2, \ldots, xm;\ \alpha 1, \alpha 2, \ldots, \alpha l)$ with a parameter α is the nonlinear function for the independent variable $x = (x1, x2, \ldots, xm) \in Rm$. For a nonlinear regression analysis, the first problem to be solved is how to obtain the best estimate of the l-dimensional parameter α.

4.1.2 Method to approximate common nonlinear mathematical models to regression model parameters

In health statistics, the most widely used non-linear mathematical models include exponential parameter, power parameter, S-growth parameter, special power parameter and exponential parameter $y = [g(x)]^{\alpha} \exp[\beta h(x)]$. For the nonlinear mathematical model obtained from the experiment, assume its data set in the m+1 -dimensional space $X - Y$ is $\{((x1, x2, \ldots, xm)i, yi) i = 1, 2, \ldots, n\}$. As per the experimental data of the nonlinear mathematical model that has not been fitted in the m+1-dimensional space $X - Y$, the theoretical prediction data set corresponding to it is $\{((x1, x2, \ldots, xm)\ i, y\ i) | i = 1, 2, \ldots, n\}$. It is difficult to directly find out the theoretical prediction value in the m+1-dimensional space $X - Y$ due to the nonlinearity of the model, so commonly alternative methods are used to make variable substitution to the nonlinear function: convert into a linear model, carry out linear regression and then revert to a nonlinear model.

Variable substitution $z = F(y)$ can be used to convert the data in the m+1-dimensional space $X - Y$ into the data in the m+1-dimensional space $X - Z$. Afterwards, the image collection of the data set in the new m+1-dimensional space X-Z is $\{((x1, x2, \ldots, xm)i, zi)\ i = 1, 2, \ldots, n\} = \{((x1, x2, \ldots, xm)\ i, F(yi)) | i = 1, 2, \ldots, n\}$. As thus, its theoretical prediction data set in the new m+1-dimensional space X-Z is $\{((x1, x2, \ldots, xm)i, zi)\ i = 1, 2, \ldots, n\} = \{((x1, x2, \ldots, xm)i, F(y\ i))\ i = 1, 2, \ldots, n\}$.

When determining the corresponding nonlinear mathematical model as per the experimental data set, some textbooks and papers usually first get the residual sum of squares in the new $m + 1$ -dimensional variable space X-Z:

$$s_1 = \sum_{i=1}^{n} (z_i - \hat{z}_i)^2 \qquad (2)$$

Then the least square method is used to determine the best estimate $\alpha = (\alpha 1, \alpha 2, \ldots, \alpha l) \in R^l$ of l-dimensional parameter α. Finally, substitute the estimate into Equation (1) to obtain the nonlinear mathematical model.

The above method is used to determine the nonlinear mathematical model, which, however, has naturally hidden a serious unnoticeable defect, which is the residual sum of squares $S1$ in the new m+1-dimensional variable space X-Z and the minimal l-dimensional parameter α does not necessarily ensure that the residual sum of squares S2 in the original m+1-dimensional variable space X-Z is minimal, where

$$s_2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \qquad (3)$$

It is this defect that has led to an error in the regression parameter of the nonlinear mathematical model, as obtained using the above method. In severe cases, it will even make the nonlinear model ineffective completely.

5 Conclusion

The combination forecasting model based on data mining can dig out more information from the original data, which is conductive to solving practical problems in different situations.According to statistic investigation, data mining has a potential huge market value and will form a new industry in China in near future, with the increase of data volume and the wide application of computer.

References

[1] Krause F L,Kaufmann U. 2007. Meta-modelling for interoperability in data mining. CIRP Annals. Vol.145, pp191-196.

[2] Trujillo J,Sergio L M. 2003. A UML based approch for modeling ETL processes in data warehouses. Conceptual Modeling 2003,Chicago,Notes in Computer Science. pp277-282.

[3]Lines J, Davis L M, Hills J, et al. A shapelet transform for time series classification[C]. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China, 2012,pp289-297.

[4]Alzghoul A , Lofstrand M., Backe B. Data stream forecasting for system fault prediction[J]. Computers and Industrial Engineering, 2012, 62(4),pp972-978.

[5]Hashemi S, Yang Y. Flexible decision tree for data stream classification in the presence of concept change, noise and missing values[J]. Data Mining and Knowledge Discovery, 2009, 19(1),pp95-131.

[6]Sandra Junier, Erik Mostert. A decision support system for the implementation of the Water Framework Directive in the Netherlands: Process, validity and useful information .Environmental Science & Policy, Volume 40, June 2014, pp49-56