Research on the Application of Data Mining Algorithm Based on Decision Tree

Song Liangong

North China University of Water Resources and Electric Power, Zhengzhou, 450045, China

Corresponding author Song Liangong

Abstract

In this paper, the author researches on the application of data mining algorithm based on decision tree. Compared with other classification methods, decision tree has the following advantages: relatively smaller calculation workload, the ease of getting apparent rules, the capability of showing important decision characteristics and higher correctness of classification, etc. However, the existing decision tree algorithm also exists a lot of shortage when being applied in practice, such as its lower computation efficiency and bigger scale of decision tree, etc. Attribute reduction algorithm ER based on the degree of dependency of attribute and post-pruning algorithm Prune based on rough set theory are proposed. Finally, the optimized decision tree algorithm is used in supplier measurement system, and its validity is verified when comparing with other algorithm.

Keywords: APPLICATION, DATA MINING, DECISION TREE

1. Introduction

With the fierce competition between companies, not only the simple way to handle and manage data from database can satisfy people's needs, but also to help manager make decision through integrating many channels of data. On the other hand, with the development of computer technology, the volume of database is up to TB level. Large-scale of data not only make user hard to operate, but also behind these data are valuable information which can be used to make decision. Data mining technology is to be developed and progressed under this background. Data Warehouse and Data Mining are new technologies for decision support, which are increasingly developed within these 10 years. Data Warehouse is to integrate comprehensive data in enterprise for Data Mining, which helps Data Mining focus on core processing. At last, they are correlate-developed. Based on the project of Large Database Pre-research, after learning data warehouse and data mining technologies, Jin's paper [1] designs data mining model under SPSS Clementine environment using C5.0 decision-tree algorithm. Because of the specialty in this project, his paper tests the correctness of data mining model using Foodmart2000 database as test data source. His paper is mainly on the design of data mining model and configuration of parameters of all nodes. After data mining, the result is analyzed to find features of members and give out some useful suggestions for development, reaching the goal of making decision based on information from database. And the correctness of results is tested using model of neural network node. Qiu's [2] paper designs and analyzes data mining model for specific theme, implementing theories and algorithms of data mining. Using Foodmart2000 database as test data source, the paper tests the feasibility and correctness of data mining model, provid-

ing theoretical and further design basement for application in the field of telecom.

With the rapid development of modern society, all kinds of information and data have the explosive growth. The huge amounts of data in medium, without the help of external tools, it is too hard to find useful information from the huge amount of data. These are far more than the human ability to understand and summarize. The emergence of Data Mining technology solved the problem very well. Data Mining can learn and analysis useful patterns and rules for the user from a large amount of data. By studying these patterns and rules, when a new sample data arises, on the basis of the existing patterns and rules we can predict the possible features for the sample. Data mining classification is one of the important steps of Data mining. Decision tree classification algorithm is a kind of widely used in Data mining, including ID3 algorithm and C4.5 algorithm, ID3 has the advantages of easy operation, but also has a preference for processing small data set, and can only deal with discrete attributes. C4.5 algorithm can make up for a lack of ID3algorithm, but also it has the shortage of the problem of incremental learning. To solve the problem of the incremental learning in the decision tree algorithm is the starting point in Li's [3] paper. In Yang's [4] paper, with the introduction of Data Mining Algorithms of the classification in detail; And then combining the classification algorithm and incremental learning technology, an incremental decision tree algorithm is proposed to solve the problem of incremental learning, and analysis the experimental data for this algorithm. In view of the common classification algorithm in Data Mining, including: the decision tree classification algorithm, k-Nearest Neighbor algorithm and the neural network classification algorithm, make a detailed introduction an description, and have carried on the comparative study with three algorithms' performance. This article selects ID3 algorithm and C4.5 algorithm for detailed research. Describing the basic steps of two algorithms in detail, including the decision tree generation and the basic steps of decision tree and listing examples to demonstrate the principle the principle of the algorithm, according to the former, combining Bayesian classification algorithm's incremental learning characteristic, an incremental decision tree algorithm is prompted, and through the analysis of experimental data, this algorithm can solve the incremental learning problem of the decision tree algorithm very well [5].

2. Data Mining Theory and Model

Data mining is the fairly significant thesis in the field of information-processing technology, which consists of many theories and technology such as database, artificial intelligence, machine learning and statistics. Classification is one of the important functions of data mining; classification algorithm based on decision tree is widely used in data mining. Compared with other classification methods, decision tree has the following advantages: relatively smaller calculation workload, the ease of getting apparent rules, the capability of showing important decision characteristics and higher correctness of classification, etc. However, the existing decision tree algorithm also exists a lot of shortage when being applied in practice, such as its lower computation efficiency and bigger scale of decision tree, etc. Therefore, it possesses significance both theoretically and factually to make further improvement in decision tree algorithm so as to enhance its capability and make it more suitable for practical application. In order to try to solve the above problems, the author of this paper makes deep research on those points. The rough set theory is introduced into decision tree classification and the method of optimizing the decision tree classification algorithm is investigated. The main work done in this paper is as follows: Firstly, this paper introduces the related technology and the theoretic basis of data mining and classification technology, and the emphasis is attached to the analysis and comparison of decision tree and post-pruning algorithms. Secondly, the decision tree algorithm is optimized in this paper in two aspects: attribute reduction and pruning. Attribute reduction algorithm ER based on the degree of dependency of attribute and post-pruning algorithm Prune based on rough set theory are proposed. Finally, the optimized decision tree algorithm is used in supplier measurement system, and its validity is verified when comparing with other algorithm.

Data mining (DM) or knowledge discover database (KDD) is to discover useful information and potential knowledge from plentiful and uncompleted and noise and fuzzy and random data which are hided and not known by people. This discovered knowledge might be used to manage information and optimize queries and make decision and control procedure and maintain database and so on. So data mining is a very valued new area of database research area, and it is a crossed subject that adopts theory and technology of database and artificial intelligent and machine learning and statistics and so on. Classification is a very important task in data mining and extensively applied to commerce at present. The destination of classification is to learn a classification function or classification model that can map a data item to a reassigned class. The researcher of machine learning and expert

system and neural biology provides a lot of classification methods. Chi's [6] paper does some research works about classification algorithm in data mining. Classification algorithm is divided to eager and lazy and total research works are based on this divide. The base technologies of classification in data mining are introduced [7]. These technologies include the procedure of classification and the preprocessing of classification data and compared and evaluated criterion of classification methods. Several of typical classification algorithms are compared which are decision-tree and k-nearest neighbor and neural network algorithm. Then the emphasis of Richard's [8] paper is induced that divide the classification to eager and lazy and the research of classification algorithm in data mining is based on this divide. A lazy decision-tree algorithm that comes from the idea of lazy classification based on model is researched on the base of the research of the traditional decision-tree. In traditional decision-tree, the concepts and advantages and disadvantages of decision-tree are presented, and the application and research situation of decision-tree are analyzed. Appling to web environment a web application used lazy decision-tree algorithm that comes from the idea of lazy based on model classification is developed. And the practical run shows this method acquired better grade. Neural network is deeply researched as representation of eager classification. Perceptron is selected. At first the creation of typical perceptron model and it's learn algorithm are introduced. Then on the base of the principal and geometrical presentation of typical perception model, the limitations of typical perceptron model are studied. This limitation is that perceptron learn algorithm can be used only when data are linear reparability. To resolve this problem, expanded perceptron models are research.4. Algebra hyper surface neutral network is a kind of expanded perceptron model. This model is an emphasis in data mining. At first the creation of this model and its geometrical presentation are introduced. Then its learning algorithm is accomplished and test's results and innovation of program are presented. At last the further aims are provide base on test's conclusion. This model is potential to resolve nonlinear reparability problems; especially it adapts to classify high-dimension data. Adaptive raise degree computer method is the innovation of research. Researches show that success rate of creating model raise after using the adaptive method. But it exists the limitation of memory for high-dimension data. So a deeply research will be continued. Figure 1 shows the web data mining process.

Based on web data mining of remote education model see figure 2, model in traditional of based on

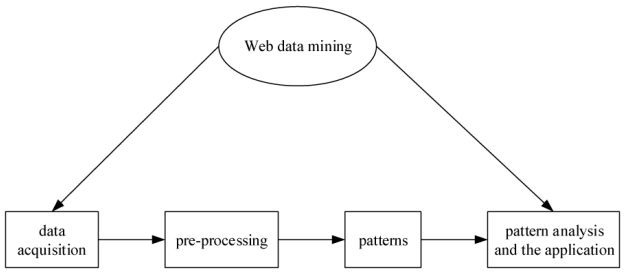

Figure 1. Web data mining process

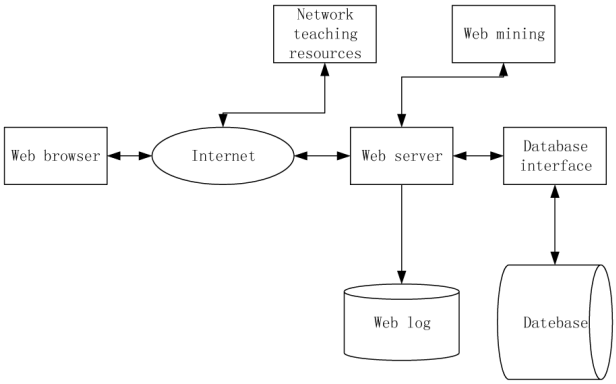

Figure 2. Distance education model based on web mining

web of remote education mode increased web mining technology through on website log and background database of integrated analysis.

3. The Algorithm

Accumulation of electronic data has taken place at an explosive rate. Undoubtedly there must be abundant latent knowledge in these electronic data of gigantic magnitude which are very important to people and traditional data analysis tools only utilize few proportion of it. Recently continually developing technic named Data Mining just can help people find latent knowledge from data. The Classification is very important method of Data Mining. Classification method can be compared and evaluated according to the following criteria: Accuracy, Speed, Robustness, Scalability, and Interpretability. Among these five criteria predictive accuracy is most important. In this paper national and international popular methods of Classification are researched and analyzed in those five aspects including classification by Decision Tree, Bayesian Classification, Classification Based on Neural Network and Classification Based on Support Vector Machine. Among these methods, Decision Tree is one of the most universal models adopted. This paper focus more on the Decision Tree, involving in the decision tree building process in all major sectors, doing a more in-depth study in the major problems of decision tree encountered on the present and future development, proposing a number of effective new ways to improve the performance of Decision Tree,

making own contribution to the further application of the Decision Tree. Attribute choosing, discretization and dimension reduction, what are the common areas of Decision Tree and other data-mining methods, not only can improve the performance of Decision Tree, but also can improve other data-mining methods. So it has positive significance to the development of data-mining technology. A novel dimension reduction algorithm is proposed in this section. A weighted binary search algorithm is proposed to discrete continuous attributes. Based on the former works, optimization and conformity is applied to the classical Decision Tree. An improvement to algorithm procedure is proposed which is shown in the following.

The containing inclusions can be simplified into the following integral equation set:

$$f(x,\omega)=f^0(x,\omega)+\int_V \mathcal{S}(x-x')(\mathrm{L}^1F(\mathrm{y}') +\rho_1\omega^2\mathbf{g}(R)\mathrm{T}_1f(\mathrm{y}')]S(\mathrm{y}')\mathrm{d}y' \quad (1)$$

In view of the following relationship

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ik_3x_3'}dx_3'=\delta(k_3) \quad (2)$$

Equation (1) can be converted into the following form:

$$f(y,\omega)=f^0(y,\omega)+ \int_s S(y-y',\omega)\mathrm{L}^1F(\mathrm{y}',\omega)\mathrm{d}y' +\rho_1\omega^2\int_s \mathbf{g}(y-y',\omega)\mathrm{J}f(\mathrm{y}',\omega)\mathrm{d}y' \quad (3)$$

After $\bar{g}(r,t)$ is obtained, $\bar{h}(r,t)$ can be easily obtained from Equation (2). So, we have:

$$\bar{g}(k,t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{-i\omega t}d\omega}{k^2+(\varepsilon-i\frac{\omega}{c})^2} =c^2\Theta(t)\frac{\sin(ckt)}{ck}e^{-\varepsilon t} \quad (4)$$

$$\bar{g}(r,t)=\frac{1}{(2\pi)^2}\int e^{-i\mathrm{kr}}\bar{g}(\mathrm{k},t)d^2k \quad (5)$$

Via Equation (4), (5) can be converted into:

$$\bar{g}(r,t)=\Theta(t)\frac{c}{(2\pi)^2}\int_0^{2\pi}d\varphi \int_0^{\infty}\sin(k[ct-kr\cos\varphi])dk \quad (6)$$

In the equation, the following property is adopted:

$$\int_0^{2\pi}\sin(kr\cos\varphi)d\varphi=0 \quad (7)$$

For defining, we normalize it

$$\int_0^{\infty}\sin k\lambda dk=\lim_{\varepsilon\to 0+}\int_0^{\infty}e^{-\varepsilon k}\sin k\lambda dk =\lim_{\varepsilon\to 0+}\mathrm{Re}\frac{1}{\lambda+i\varepsilon} \quad (8)$$

Thus, (6) can be represented into:

$$\bar{g}(r,t)=-\frac{c\Theta(t)}{(2\pi)^2}\mathrm{Re}\int_0^{2\pi}\frac{d\varphi}{r\cos\varphi-ct+i\varepsilon} \quad (9)$$

Thus, Equation (9) can be represented as:

$$\bar{g}(r,t)=-\frac{c\Theta(t)}{(2\pi)^2}\frac{2}{r}\mathrm{Re} \int_0^{2\pi}\frac{e^{i\varphi}d\varphi}{e^{2i\varphi}-2\cosh\phi e^{i\varphi}+1} =-\frac{c\Theta(t)}{(2\pi)^2}\mathrm{Re}\frac{2}{ir}\oint_{|s|=1}\frac{ds}{s^2-2\cosh\phi s+1} \quad (10)$$

Within the unit cycle to solve (10), the following can be obtained:

$$\bar{g}(r,t)=-\frac{c\Theta(t)}{2\pi}\frac{2}{r}\mathrm{Re}\frac{1}{\sinh\phi} \quad (11)$$

$$\bar{g}(r,t)=-\frac{1}{2\pi}\frac{\Theta(t-\frac{r}{c})}{\sqrt{t^2-(\frac{r}{c})^2}} \quad (12)$$

$\bar{h}$ expression can be obtained:

$$\bar{h}(r,t)=\frac{\Theta(t-\frac{r}{c})}{2\pi} \left\{t\ln\left(\frac{ct}{r}+\sqrt{\frac{c^2t^2}{r^2}-1}-\sqrt{t^2-\frac{r^2}{c^2}}\right)\right\} \quad (13)$$

Thus, $\bar{g}(r,\omega)$ must be determined, that is:

$$\bar{g}(r,\omega)=\int_{-\infty}^{\infty}e^{i\omega t}\bar{g}(r,t)dt \quad (14)$$

Put (13) into (14) to obtain:

$$\bar{g}(r,\omega)=\frac{1}{2\pi}\int_{r/c}^{\infty}\frac{e^{i\omega t}dt}{\sqrt{t^2-\frac{r^2}{c^2}}} \quad (15)$$

$$\bar{g}(r,\omega)=\frac{1}{2\pi}\int_0^{\infty}e^{i\frac{\omega r}{c}\cosh\phi}d\phi \quad (16)$$

$$H_0^1(z)=\frac{2}{\pi i}\int_0^{\infty}e^{iz\cosh\phi}d\phi \quad (17)$$

As per (17), $\bar{g}(r,\omega)$ of Equation (18) can be obtained:

$$\bar{g}(r,\omega)=\frac{i}{4}H_0^1(\frac{\omega r}{c}) \quad (18)$$

The function component defined for Fourier transform can be obtained in the following equation (19)-(21).

$$G_{ik}(r,\omega)=\frac{i}{4\rho_0\omega^2}\{\theta_{ik}\beta^2H_0^1(\beta r)$$

$$-\frac{\partial^2}{\partial y_i\partial y_k}[H_0^1(qr)]_\beta^\alpha+m_im_k\beta_\perp^2H_0^1(\beta_\perp r)\}$$

$$\gamma_i(r,\omega)=\frac{i}{4\rho_0\omega^2}(\frac{e_{15}^0}{\eta_{11}^0})\beta_\perp^2H_0^1(\beta_\perp r)m_i \tag{19}$$

$$g(r,\omega)=\frac{1}{2\pi\eta_{11}^0}\ln r+\frac{i}{4\rho_0\omega^2}(\frac{e_{15}^0}{\eta_{11}^0})^2\beta_\perp^2H_0^1(\beta_\perp r)$$

In which,

$$[f(qr)]_\beta^\alpha=f(\alpha r)-f(\beta r),\ r=|y| \tag{20}$$

$$G_{ik}(r,t)=\{\frac{\theta_{ik}}{C_{66}^0}\bar{g}_2(r,t)+\frac{1}{\rho_0}\frac{\partial^2}{\partial y_i\partial y_k}$$

$$[\bar{h}_2(r,t)-\bar{h}_1(r,t)]_\beta^\alpha+\frac{m_im_k}{C_{44}'}\bar{g}_3(r,t)\}$$

$$\gamma_i(r,t)=\frac{e_{15}^0}{\eta_{11}^0C_{44}'}m_i\bar{g}_3(r,t) \tag{21}$$

$$g(r,t)=\frac{\delta(t)}{2\pi\eta_{11}^0}\ln r+\frac{(e_{15}^0)^2}{(\eta_{11}^0)^2C_{44}'}\bar{g}_3(r,t)$$

In which, $\bar{g}_i$ represents output of Equation (22)

$$\bar{g}_i(r,t)=\frac{1}{2\pi}\frac{\Theta(t-\frac{r}{c_i})}{\sqrt{t^2-(\frac{r}{c_i})^2}} \tag{22}$$

Meanwhile, it also represents output for equation (5)

$$\bar{h}_i(r,t)=\frac{\Theta(t-\frac{r}{c_i})}{2\pi}$$

$$\left\{t\ln\left(\frac{c_it}{r}+\sqrt{\frac{c_i^2t^2}{r^2}-1}\right)-\sqrt{t^2-\frac{r^2}{c_i^2}}\right\} \tag{23}$$

It can be represented as:

$$c_1=\sqrt{\frac{C_{11}^0}{\rho_0}},\ c_2=\sqrt{\frac{C_{66}^0}{\rho_0}},\ c_3=\sqrt{\frac{C_{44}^0}{\rho_0}},$$

$$C_{44c}'=C_{44}^0+\frac{(e_{15}^0)^2}{\eta_{11}^0} \tag{24}$$

It can be further represented as:

$$u_i(y)=u_i^0(y)+$$

$$\int_s[\Psi_{imn}(R)\varepsilon_{mn}(y')-\psi_{im}(R)E_m(y')+\rho_1\omega^2G_{ik}(R)u_k(y')]dy' \tag{25}$$

In which,

$$\Psi_{imn}(R)=G_{ik,l}(R)C_{klmn}^1+\gamma_{i,k}(R)e_{kmn}^{1T},$$

$$\psi_{im}(R)=G_{ik,l}(R)e_{mkl}^1-\gamma_{i,k}(R)\eta_{km}^1, \tag{26}$$

$$\Phi_{mn}(R)=\gamma_{k,l}(R)C_{klmn}^1+g_{,k}(R)e_{kmn}^{1T},$$

$$\phi_m(R)=\gamma_{k,l}(R)e_{mkl}^1-g_{,k}(R)\eta_{km}^1,$$

$$R=|y-y'| \tag{27}$$

Thus, the following can be obtained:

$$u_i(y)=u_i^0(y)+u_i^s(y),\ \varphi(y)=\varphi^0(y)+\varphi_i^s(y) \tag{28}$$

Conclusions

In this paper, the author researches on the application of data mining algorithm based on decision tree. Compared with other classification methods, decision tree has the following advantages: relatively smaller calculation workload, the ease of getting apparent rules, the capability of showing important decision characteristics and higher correctness of classification, etc.

The Classification is very important method of Data Mining. Classification method can be compared and evaluated according to the following criteria: Accuracy, Speed, Robustness, Scalability, and Interpretability. Among these five criteria predictive accuracy is most important. In this paper national and international popular methods of Classification are researched and analyzed in those five aspects including classification by Decision Tree, Bayesian Classification, Classification Based on Neural Network and Classification Based on Support Vector Machine.

The huge amounts of data in medium, without the help of external tools, it is too hard to find useful information from the huge amount of data. These are far more than the human ability to understand and summarize. The emergence of Data Mining technology solved the problem very well. Data Mining can learn and analysis useful patterns and rules for the user from a large amount of data. By studying these patterns and rules, when a new sample data arises, on the basis of the existing patterns and rules we can predict the possible features for the sample. Data mining classification is one of the important steps of Data mining. However, the existing decision tree algorithm also exists a lot of shortage when being ap-

plied in practice, such as its lower computation efficiency and bigger scale of decision tree, etc. Attribute reduction algorithm ER based on the degree of dependency of attribute and post-pruning algorithm Prune based on rough set theory are proposed. Finally, the optimized decision tree algorithm is used in supplier measurement system, and its validity is verified when comparing with other algorithm.

References

1. Xiaoling Jin, Yong Wang, Zhilong Huang, Mario Di Paola. Constructing transient response probability density of non-linear system through complex fractional moments. International Journal of Non-Linear Mechanics, 2014, pp. 65-76.
2. Xiang QIU, HUANG Yong-xiang, Quan ZHOU, Chao SUN. Scaling of maximum probability density function of velocity increments in turbulent Rayleigh-Bénard convection. Journal of Hydrodynamics, Ser.B, 2014, pp. 263-278.
3. Bo Li, Fu-Wen Pang. Improved cardinalized probability hypothesis density filtering algorithm. Applied Soft Computing Journal, 2014, pp. 24-39.
4. Qingshan Yang, Yuji Tian. A model of probability density function of non-Gaussian wind pressure with multiple samples. Journal of Wind Engineering & Industrial Aerodynamics, 2014, pp. 541-549.
5. D. McDonagha, A. Brusebergb, C. Haslamc. Visual product evaluation: exploring users' emotional relationships with products. Applied Ergonomics, 2002, 33, p.p.231-240.
6. Kaikai Chi, Yi-hua Zhu, Xiaohong Jiang, Xianzhong Tian. Practical throughput analysis for two-hop wireless network coding. Computer Networks, 2013, pp. 233-256.
7. Luiz Filipe M. Vieira, Mario Gerla, Archan Misra. Fundamental limits on end-to-end throughput of network coding in multi-rate and multicast wireless networks. Computer Networks, 2013, pp. 5717-5727.
8. Richard R. Young, Peter F. Swan, Evelyn A. Thomchick, Kusumal Ruamsook. Extending landed cost models to improve offshore sourcing decisions. International Journal of Physical Distribution & Logistics Management, 2009, pp. 394-402.