Credit Risk Assessment of Peer-to-Peer Lending Borrower Utilizing BP Neural Network

Zhengnan Yuan[1], Zihao Wang[2], and He Xu[2,3](✉)

[1] Glasgow College, UESTC,
University of Electronic Science and Technology of China, Chengdu, China
916260629@qq.com
[2] School of Computer,
Nanjing University of Posts and Telecommunications, Nanjing, China
1390424429@qq.com, xuhe@njupt.edu.cn
[3] Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, China

Abstract. This paper proposes an innovated approach of risk assessment of borrowers based on the BP neutral network model. Specifically, firstly, referring to the empirical data published by the website 'peer-to-peer lender' and the indicators of personal credit risk assessment from commercial bank is an efficient method to pick several valid values through data processing, classification and quantification, then the final modeling indicators are selected by information gain technology. Secondly, the new credit risk assessment model is formed after training the modeling indicators. Meanwhile, several strings of collected testing data would be substituted to find out the default rates which are supposed to be compared with the practical ones on the website and the calculated ones from existing credit risk assessment evaluating models. Last but not the least, the effect of this new method is evaluated.

1 Introduction

P2P (Peer-to-peer) Website which is under an admirable escalation with the promotion from the comprehensive regulatory and the new international marketing environment has been operating for over a decade [1–3]. However, as a result of the limitation in scale, comparing with traditional commercial banks, peer-to-peer lending companies demonstrate a weak controllability when facing the varying risks, ranging from credit risk, legal risk, regulatory risk, information asymmetry, investment risk, discipline risk, settlement risk and information security risk. Among those the most serious problem is the credit risk – most of the platforms have not polish up their model of risk assessment whereas present typical credit assessment model of lending has innovated into a totally different state. Simply applying those models that are merely suitable for classical traditional finance model could not be practical.

The most innovative part of this study is utilizing BP neutral network, information gain technology and introduction of values to evaluate the credit risk assessment of P2P

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_3

lending and getting a desired result, conclusion and testing results. On top of that, some expected advancements which root in the conclusion of this study on personal credit assessment of peer-to-peer lending borrower based on BP neural network are proposed at this paper.

The internet financial is becoming different from not only indirect financing of traditional commercial banks, but also the capital market whose indirect fund is raised by direct fund new financial model with the improvement of modern information technology especially the internet. Starting from 2007, P2P lending came to China and gradually becomes the delegate of internet financial models [4–7]. P2P network lending is designed to match the requirement of individual borrowers and the loans of small or medium-sized entrepreneurs. As the intermediary platform, P2P network platform allow the individuals and small or medium-sized entrepreneurs that have idle funds or are willing to loan to publish the information of loan with rate and due date selected by themselves. As a result, more deals could be done by this kind of method.

2 BP Neutral Network

BP neutral network (Background Propagation), also called error back propagation network, was firstly introduced by Werbos in 1974. In 1985, Rumelhart and other scholars did effort to develop the theory and propose clear and strict algorithms [8]. BP algorithms is applied to forward network. It applies the training forms which the tutors' help is involved, and it can also provide both the input and the output vector product simultaneously. By using the back propagation learning algorithms and adjusting the link weight of network, the network output is expected to be approximate to the expected output to the greatest extent under the condition of least mean square error. The progress of backward learning consists of forward and backward propagation. Specifically, in the process of forward propagation, the input information transfers to output layer after the testing of hidden neutron, if the output layer could not receive the output as expected, then the information will transfer to the backward propagation process where the error of the actual output compared to the expected one will be sent back in the former connected channel [9]. Eventually, through modifying authority of the connecting of each layer neurons, the errors can be reduced, and then it can transfer to the forward propagation process where a recycle is formed until the error is less than a given value.

3 Application Flow Chart

Firstly, being a new product of the modern information society, P2P lending has not established efficient systems that is related to the information risk evaluation mechanism so far, and the evaluation for the information risk of borrower is not impeccable. However, BP neutral network has the characteristics of self-adjusting, high self-study and high flexibility which can adjust itself merely according to the variations of the environment, then find regulations for large amount of data and provide relatively correct inference results based on those regulations. Thus, BP neutral network has strong practical feasibility on the P2P lending's defects which includes uncertainty of information, lacking of efficient systems that is related to the information risk evaluation mechanism and the evaluation for the information risk of borrower. In addition, BP neutral network can display the professors' knowledge, experience and thoughts, thus it can get rid of the subjective evaluation as much as possible. Then it is obvious that the credit evaluation can be more precise. In the end, BP neutral network model is a nonlinear modeling process which is not necessary to learn the nonlinear relationship between data. Technically, this indicates that it can effectively overcomes the difficulties of choosing the suitable model functions in the traditional modeling process, and it can establish the modeling speedily. As a result, it can be applied in various fields. The algorithm flow chart for using BP is shown in Fig. 1.

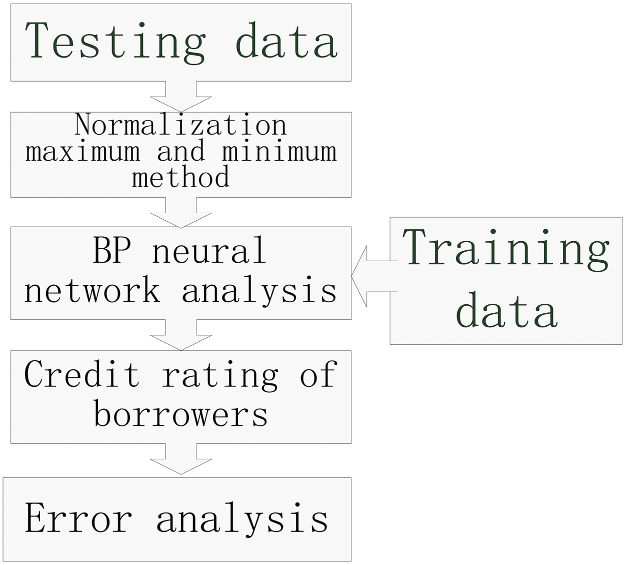

Testing data
Normalization maximum and minimum method
BP neural network analysis
Training data
Credit rating of borrowers
Error analysis

Fig. 1. Algorithm flow chart

4 Model Construction

4.1 Target Selection

P2P network lending platform generally requires the borrower to provide personal information including the identity, occupation, property and other personal basic conditions. After this state is complicated, the information borrowers provide is judged through the field of authentication, video authentication, and so on, to ensure the authenticity of the information. According to the certification, followed by the assessment of the credit rating of the borrowers, the information and credit rating results will be published on the website as a reference for lenders. Therefore, according to the characteristics of P2P lending on the website, the data in Renren Loaning Website and the selection principle of the traditional commercial bank personal credit rating index, the basic information of the borrower is concluded into five aspects: demographic characteristics, occupation status, income and property, credit history operation and certification. Considering of the personal credit rating of commercial banks which is combined with the characteristics of P2P network platform, various indicators of the importance of the credit rating and five aspects of credit evaluation index are quantitative. The reasons for the selection of the index and the value are as follows:

Demographic characteristics: demographic characteristics include age, marital status and educational level. For age, there is a significant difference in default rates among borrowers of different ages. Generally, the default rate of individuals from 35 to 50 years-old who own a stable job and in satisfied economic condition is low. On top of that, for 26- to 35-year-old borrowers, although their income may experience an increase, the pressure comes from the families are serious problem of rising the default rate to a general level. Borrowers of whose ages are below 25, their low income, the lack of mercury spending habits and mostly no saving capabilities all contribute to the high risk of defaulting. Those older than 50 years old, whose level of income begin to decline, are more likely to have sudden consumption. As a result, the default risk is relatively large. Apart from that, marital status is also a key to this, married borrowers are more reliable and divorces or unmarried borrowers may be in low credit status. Lastly, the education is also essential, overall, the higher the education level is, the lower the probability of occurrence of default is.

Occupational status: Occupational status, ranging from unit type, type of jobs and years of services. As for unit type, generally speaking, people who work for the government could have a more stable source of income and have less likelihood of default. Therefore, they are of the highest value. The larger the firm sizes, the more stable income level they have. As a result, the default rate is smaller. For post type, higher post has higher level of earnings with small risk of default. Moreover, for work experience, the longer years of working leads to a more reliable income levels. Those individuals also seem to have lower risk of default.

Income property: First of all, the high the income is, the smaller the default is. Secondly, as for properties, especially that of China, housing conditions normally represent individual economic capacity, so there are existing a smaller risk of default than that of non-real state. Last but not the least, cars can also represent the economic capacity of people, that is to say, the car owners are less likely to default.

Credit history: Credit history is mainly reflected by the number of successful repayment (i.e. Successful repayment times: The more number of successful repayment, the better the credit inertia is and the less likely to violate rules) and the number of overdue repayment (i.e. overdue repayment times: the more number of overdue repayment, the worse the history of credit history is and the easier to violate rules).

Operation certification: The more kinds of authentication, the more reliable the information is, then the more complete and the smaller the possibility of default is. According to the standard of the personal credit rating, combined with the characteristics of P2P network platform, the higher index value and the higher credit rating, the smaller the possibility of default is.

4.2 Data Processing

This paper extracts transaction data from a number of online trading platforms (i.e. a total of 14 k borrower information) as a sample of P2P personal borrowers' credit risk assessment. Then those data would be transformed into quantitative data according to the personal credit indicators in qualitative indicator [10].

In general, the input sample values of the neural network are required to be normalized. In this paper, the maximum and minimum method is utilized to normalize the quantitative data of personal credit indicators, that is to say, using the formulas below to normalize. Maximum and minimum method is a kind of linear transformations that will not cause too much loss according to formula (1). And our code for BP network to process the data is shown as the following.

$$\mathbf{u}_i = \frac{\boldsymbol{u} - \boldsymbol{min}(\boldsymbol{u})}{\boldsymbol{max}(\boldsymbol{u}) - \boldsymbol{min}(\boldsymbol{u})} \tag{1}$$

```
pseudo-code:
Program prepare data for BP network (Input)
{Gain main information}Repeat
i=42:1:50:[data,text]=xlsread(strcat(''dataintegration/traini
ng data/platform',num2str(i),'.csv'));
count=size(text,1);
i=1:10303,strcmp(listO(i),'installment')==1:n_listO(i)=1;
else : strcmp(listO(i),'one-off payment')==1: n_listO(i)=2;
else : n_listO(i)=3;end;
i=1:10303, strcmp(listQ(i),'male')==1:n_listO(i)=1;
else : n_listO(i)=2;end;
i=1:10303,strcmp(listS(i),'underguaduate')==1||strcmp(listS(i
),'graduate')==1: n_listS(i)=1;
else : n_listO(i)=2;end;
i=1:10303, strcmp(listT(i),'married')==1: n_listS(i)=1;
else : strcmp(listT(i),'unmarried')==2;
else : n_listO(i)=3;end;
i=1:10303, strcmp(listW(i),'\N')==1: n_listW(i)=1;
else : n_listW(i)=2;end;
i=1:10303, strcmp(listX(i),'yse')==1: n_listX(i)=1;
else : n_listX(i)=2;end;
i=1:10303, strcmp(listY(i),'yse')==1: n_listY(i)=1;
else : n_listY(i)=2;end;
Program training data for BP network (Input)
{Process for F data including yuan or RMB}
y1=strfind(S, 'yuan');r1=strfind(S, 'rmb')；
r2=strfind(S, 'RMB');Repeat
if there exists any these key word: transform data to lisfF
{Process for J data including loan rate}
listJ=data(:,10);
money_rate: aJ=data(:,10); if listJ(2)<=1 listJ=listJ*100;
listO=text(:,14);%repay_type
listQ=text(:,16);%borrower_sex
listR=data(:,18);%borrower_age
if there exists any these key word: get digital information
in the string one by one and transform data to lisfF
{Input data}i=length(aF); b=text{i,5};
A=isstrprop(b,'digit');B=b(A); C=str2num(B);listF(i)=C;
listJ=[listJ;data(:,10)];if listJ(2)<=1;listJ=listJ*100;
{Input list information}
i=1:length(listO):S={'a'};S=listO(i);
{Input the kinds of loan} n_listO(i)=n(n=1~4);
{Input gender} n_listQ(i)=n(n=1~3);
{Input education} n_listS(i)=n(n=1~5);
{Input marriage state}n_listW(i)=n(n=1~2);n_listT(i)=n(n=1~4);
{Input house information state} n_listX(i)=n(n=1~3);
{Input car information} n_listY(i)=n(n=1~3);end
If i<=27553:state(i)=1;
Else  state(i)=-1; end;
```

5 Model Processing

5.1 Model Description

In this paper, the personal credit risk assessment process of P2P internet lending platform is simulated by the three-layer neural network [11]. Input layer mode number is 11.

The output layer is the credit rating of the individual borrower of the P2P platform which refers to the classification of platforms. The number of nodes in the output layer is 1 and the selected values are 10, 8, 6, 4 and 2, corresponding to the five credit levels, respectively. Specifically, the highest credit rating is 10 (i.e. the least likely to default), a minimum of 2 credit rating value of 1 (i.e. the lowest level which is the most likely breach of contract and cannot repay in time). The approximate range of the number of nodes in the hidden layer is firstly determined by the golden section method, and then the optimal number of nodes in hidden layers is determined through experiment.

5.2 Model Simulation

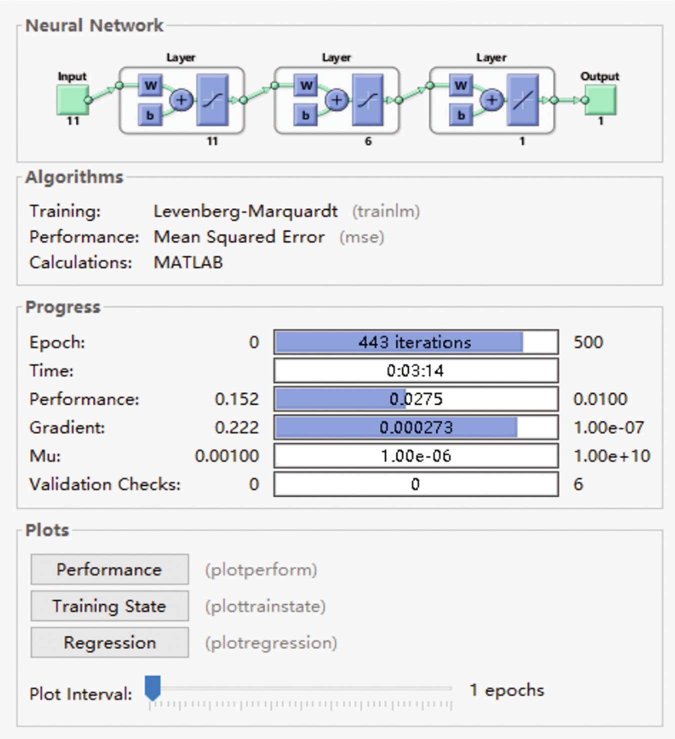
Before the simulation, 14000 sets of data from 10 different platforms are simulated and integrated as training data including 2000 of defaulted and 12000 of clean loan records. Training function is used to build BP neural network with epoch set to 500, mean squared error to 0.001 and the number of hidden layer to 5. All those are shown in Fig. 2.

This program is aimed to evaluate the risk of P2P station. There are 40 sets of training data including 16 sets of normal data which are provided by working stations and 24 sets of abnormal data which are provided by bankrupt stations. And the given 10 sets of predicting data are used to evaluate the risk.

Each station contains a number of records of loan, and the risk of P2P station is relevant to the evaluation of every record, so we decide to evaluate the risk of P2P station by evaluating the risk of each record of loan. It is obvious that the station is considered as a bankrupt station when the number of the risky records in this station exceed the threshold.

Referring to the information provided by experience and cogitate the weight of each property, 11 properties are chosen as input:

column F ITEM_AMOUNT: the amount of the loan money, it is a key property.
column J MONEY_RATE: the rate of this loan, it is a key property.
column O REPAYTYPE: the repay type of this loan, it is a key property.
column Q BORROWER_SEX: the gender of the borrower, it is borrower's personal information.



Neural Network

Algorithms

Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MATLAB

Progress

Epoch:	0	443 iterations	500
Time:		0:03:14	
Performance:	0.152	0.0275	0.0100
Gradient:	0.222	0.000273	1.00e-07
Mu:	0.00100	1.00e-06	1.00e+10
Validation Checks:	0	0	6

Plots

Performance (plotperform)
Training State (plottrainstate)
Regression (plotregression)

Plot Interval: 1 epochs

Fig. 2. Settings before training

column R BORROWER_AGE: the age of the borrower, it is borrower's personal information.
column S BORROWER_EDUCATION: the degree of the borrower, it is borrower's personal information.
column T BORROWER_MARRIAGE: the marriage of the borrower, it is borrower's personal information.
column W BORROWER_INCOME: the income of the borrower, it is borrower's personal information.
column X BORROWER_HOUSE: the house of the borrower, it is borrower's personal information.
column Y BORROWER_CAR: the car of the borrower, it is borrower's personal information.
column Z REWARD: the reward of this loan.

6 Data Analysis

Since the information of the P2P network platform is entered by the borrowers and is not mandatory, it is possible that some information is missed from the borrowers or the borrowers intentionally conceal the information which causes some mistakes in the information. The results are shown in Fig. 3. Therefore, when an individual borrower evaluates a credit risk, there are some missing or invalid information. One of the characteristics of the BP neural network model is offering a more accurate grading result by the training result in the case of partial data missing. This paper excluded some indicators, ranging from state condition, passenger vehicles, however, the output of the model and the output of the target are the same. In the absence of unit type, job type and income range, the difference between the model output and the target output is large. On top of that, if the number of successful borrowing, out-of-date repay and lack of confirming information would cause a huge gap between the model output and the target output. However, there is no reversal result which means that the borrower with low credit risk will not be regarded as a borrower that has high credit risk. In a nutshell, this data can still be regarded as the basis for credit risk evaluation of borrowers in P2P network credit. Thus, in the absence of fuzzy information, BP neural network model still has the ability of P2P network credit borrower credit prediction and the assessment accuracy rate is still high. Our data code is shown in the following.

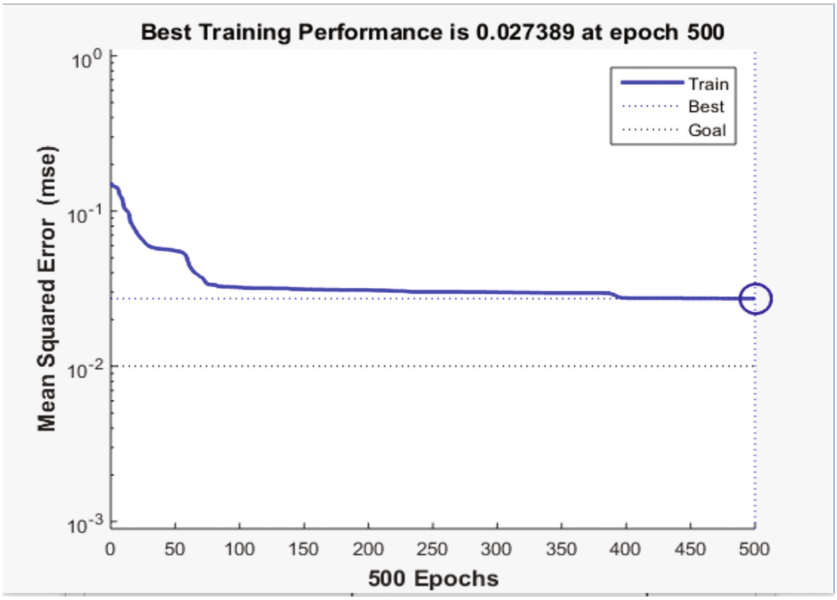


Fig. 3. The accuracy of the valuation of data from 10 different platforms is 80%.

```
pseudo-code:
Program Credit Risk Assessment (Output)
{Load train data and test data with normalization};
P=[listF';listJ';n_listO;n_listQ;listR';n_listS;n_listT;n_lis
tW;n_listX;n_listY;listZ'];
[p1,minp,maxp,t1,mint,maxt]=premnmx(P,state);
a=[listF';listJ';n_listO;n_listQ;listR';n_listS;n_listT;n_lis
tW;n_listX;n_listY;listZ'];
normalization:aa=tramnmx(a,minp,maxp);
output: b=sim(net,a);
predicting data: c=postmnmx(b,mint,maxt);
{Build and set network}
net=newff(minmax(P),[11,6,1],{'tansig','tansig','purelin'},'t
rainlm');
net.trainParam.epochs =500;
net.trainParam.goal=0.01;
net.trainParam.lr = 0.05;
  repeat
i=1:10: cnt(i)=0;
i=1:10303, i<634,  c(i)>0:cnt(1)=cnt(1)+1;
i=1:10303, 634<i<1249, c(i)>0:cnt(2)=cnt(2)+1;
i=1:10303, 1249<i<1266, c(i)>0:cnt(3)=cnt(3)+1;
i=1:10303, 1266<i<3915, c(i)>0:cnt(4)=cnt(4)+1;
i=1:10303, 3915<i<4124, c(i)>0:cnt(5)=cnt(5)+1;
i=1:10303, 4124<i<4752, c(i)>0:cnt(6)=cnt(6)+1;
i=1:10303, 4752<i<5285, c(i)>0:cnt(7)=cnt(7)+1;
i=1:10303, 5285<i<7809, c(i)>0:cnt(8)=cnt(8)+1;
i=1:10303, 7809<i<9036, c(i)>0:cnt(9)=cnt(9)+1;
i=1:10303, 9036<i<10303,c(i)>0:cnt(10)=cnt(10)+1;
number=[300 300 10 1300 400 340 270 1400 800 800];
i=1:10, cnt(i)>number(i)∶pt(i)=1;
i=1:10, cnt(i)<number(i)∶pt(i)=0; end;
```

7 Conclusion

The credit risk assessment model of credit borrower of P2P network based on the BP neural network in this study works well since the credit risk of the personal borrower can still be measured accurately even though some information is missing or ambiguous. Specifically, this study that has a certain applicability is expected to be popularized and used. The reason why the BP neural network credit risk assessment model in this study has a reliable ability of evaluation is that the BP neural network itself is good at the discovery of the knowledge and extraction of characteristic values which is suitable for the credit assessment. Overall, the results of the whole experiment demonstrate that BP neural network is a desirable selection for the credit risk assessment of individual borrower in the P2P network. In a nutshell, according to the conclusion and results of the experiment above, this paper would like to propose the countermeasures and suggestions to not only improve the credit risk evaluation of P2P borrowers but also facilitate a healthy processing of P2P platform:

Firstly, strengthen the information authentication of the P2P network lending platform to ensure the accuracy of personal information. As is known to all, the credit rating of the borrower is based on the information provided by the borrowers, so the authenticity and accuracy of the information are the key to the rating system. To improve this, the website is supposed to carry out real-time authentication or certification of the information to avoid misleading information which could do harm to the profit of the lenders.

Secondly, increase the disclosure of P2P network lending platform information. Since more personal information could help the lenders have a more comprehensive understanding for the borrowers and so as the internet to adjust the credit rating of borrowers.

Thirdly, P2P network lending platform is expected to disclosure the overdue repayment list on time. On the one hand, the number of overdue repayments is critical for the credit rating. The rating could be adjusted properly through the on-time disclosure. If this aim is achieved, the lender is able to know the real condition of the borrowers. On the other hand, under this invisible pressure from the disclosure, the borrower could repay the money on time and pay attention to the credit.

Acknowledgments. This work is financially supported by the National Natural Science Foundation of P.R. China (No. 61373017, No. 61572260, No. 61572261, No. 61672296, No. 61602261), the Natural Science Foundation of Jiangsu Province (No. BK20140886, No. BK20140888), Scientific & Technological Support Project of Jiangsu Province (No. BE2015702, BE2016185, No. BE2016777), Natural Science Key Fund for Colleges and Universities in Jiangsu Province (No. 12KJA520002), China Postdoctoral Science Foundation (No. 2014M551636, No. 2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 1302090B, No. 1401005B), Jiangsu Postgraduate Scientific Research and Innovation Projects (SJLX16_0326), Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks (WSNLBZY201509), NUPTSF (Grant No. NY214060, No. NY214061) and the STITP projects of NUPT (No. XZD2016032 and No. XYB2016532).

References

1. Chen, Y.F., Wu, C.J.: Influence of website design on consumer emotion and purchase intention in travel websites. Int. J. Technol. Hum. Interact. (IJTHI) **12**(4), 15–29 (2016)
2. Sula, A., Spaho, E., Matsuo, K., et al.: A new system for supporting children with autism spectrum disorder based on IoT and P2P technology. Int. J. Space-Based Situated Comput. **4** (1), 55–64 (2014)
3. Di Stefano, A., Morana, G., Zito, D.: QoS-aware services composition in P2PGrid environments. Int. J. Grid Util. Comput. **2**(2), 139–147 (2011)
4. Sawamura, S., Barolli, A., Aikebaier, A., et al.: Design and evaluation of algorithms for obtaining objective trustworthiness on acquaintances in P2P overlay networks. Int. J. Grid Util. Comput. **2**(3), 196–203 (2011)
5. Takeda, A., Oide, T., Takahashi, A.: Simple dynamic load balancing mechanism for structured P2P network and its evaluation. Int. J. Grid Util. Comput. **3**(2–3), 126–135 (2012)
6. Eftychiou, A., Vrusias, B., Antonopoulos, N.: A dynamically semantic platform for efficient information retrieval in P2P networks. Int. J. Grid Util. Comput. **3**(4), 271–283 (2012)

7. Higashino, M., Hayakawa, T., Takahashi, K., et al.: Management of streaming multimedia content using mobile agent technology on pure P2P-based distributed e-learning system. Int. J. Grid Util. Comput. **5**(3), 198–204 (2014)
8. Holyoak, K.J.: Parallel distributed processing: explorations in the microstructure of cognition. Science **236**, 992–997 (1987)
9. Rochester, N., Holland, J., Haibt, L., et al.: Tests on a cell assembly theory of the action of the brain, using a large digital computer. IRE Trans. Inf. Theory **2**(3), 80–93 (1956)
10. Hoskins, J.C., Himmelblau, D.M.: Process control via artificial neural networks and reinforcement learning. Comput. Chem. Eng. **16**(4), 241–251 (1992)
11. Ciresan, D., Giusti, A., Gambardella, L.M., et al.: Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems, pp. 2843–2851 (2012)