MathorCup Global Mathematical Modeling Challenge Undertaking

We have carefully read the Mathorcup Global Mathematical Modeling Challenge rules.

We fully understand that after the start of the challenge， it’s forbidden for the team members to communicate in any way (including telephone, email, online consulting, etc.) with anyone outside the team (including the instructor) to study or discuss issues related to the challenge issues.

We know that copying other people's results is a violation of contest rules, if the reference to the results of other people or other publicly available information (including information found on the Internet), must be in accordance with the provisions of the presentation of references cited in the text and references explicitly at listed.

We solemnly pledge to strictly comply with competition rules to ensure fair competition. Any violation of competition rules of behavior, we will be dealt with severely.

Problem Number（A/B/C）： B

Team Number： 10352

Can be published： YES (YES/NO)

Member： 1. Yunfei Xing

2. Lina Zhang

3. Yingzhao Song

Advisor： Rongchuan Liao

Date： May 28th 2014

Books Recommendation

Abstract

With the development of the technology of information and internet, we came to an information overloaded time. So does the book market. On one hand, it may be a rather difficult choice for the readers to select their preferred and high quality books; while on the other hand it's also of great difficulty to recommend their books to the suitable people for the writers or the book sellers themselves. Based on the statistics provided, we deeply explore the inner relations among the different kinds of data in order to build effective models to rank the books, hence recommending them to the target readers.

As for question one, our main task is to pre-process the given data to find out the possible relations among them. We adopt the association rule to the massive statistics and then come to the dimension reduction. After this process, we came to the conclusion that the heat of the book tags confine to the heavy-tailed distribution. Therefore, we set up a mapping table of the statistics, and reached the highly related influencing factors via using the missing value handling method to fill in the completed matrixes. In the end, we came to the following two major factors which may influence the readers' remarks of the books, which are the reading interest of the users and the prevailing extent of the books.

In the second question, we suppose that the coding ID of the given books are coded by Dewey Decimal Number. On the basis that we have pre-processed the given statistics, we randomly select the book types of 60000 users as the calibration of the neural network, taking the corresponding heat of the book remarks as the input terminal. In this context, we can regard the relationship between the remarks and the influence factors as a black box, which means this question can be seen as a black-box question. We then use the BP network to train the input and output data, which can fully take advantage of the nonlinear system of the BP network, hence making the prediction much more precise, because we finally adopt the well-trained network to undergo the prediction, and the final result can be seen in the body of the paper.

While, as to the question three, we use the collaborative filtering method based on the clustering process. In this method, readers who enjoy the similar interests will be clustered together, and then the neighbors who share the most similar appetites will be selected out. And then we can obtain the invisible message of the uses via their neighbors to select the top 3 books as the final recommendations. The detailed recommendation book lists are adhere to the end of the paper.

Key Word: Massive Data Mining; Heavy-tailed Distribution; BP Network; Clustering; Collaborative Filtering Analysis

一、 Restatement

1.1Background

As the prevailing of the internet, it becomes much more popular for the book industry, hence making it difficult to select a god-wants-to book for the readers, because they face a much bigger book market and the choices they can make are gradually massive. To solve this problem and pace up the requirement of the time itself, personal recommendation just right comes to the being. In this method, readers' preferences can be found in their social behavior statistics and their history data. By carefully analyzing these massive data, it finally comes to a recommendation list, and in this list, users can easily find their appetites, which helps the users a lot. However, in current china, researches on the book remarks are still on the level of qualitative analysis, both ether practical and theoretical analyses are not well advanced, hence making it possible to carry on a detailed analysis on the users' information and their historical behavior data to set up a much more precise book recommendation system.

1.2 Propose the problem

According to the given statistics and the requirements, we can come to the following three detailed questions:

1. To explore the inner relations among the massive statistics and find out the possible relationships between the remarks and the related data, in order to conclude the influence factors of the users' remark,
2. To build the proper prediction model to remark on the unread books to make a much more precise prediction to the users,
3. And then, to recommend the target books to the corresponding users via analyzing the social statistics of the users by collaborative filtering method.

二、 Analysis of the Problem

2.1 About Question 1

According to the question itself, what should be mostly focused is to find out the inner relations among the massive statistics. In this context, the key point lies to how to obtain the influence factors from hundreds of thousands of data. Therefore, the foremost thing in question 1 is to deeply explore the provided statistics and reach the final major factors which may affect the remarks of the users.

2.2 About Question 2

Based on the first question, what should be solved in the question 2 is to detailed analysis the influence factors obtained from question 1and take fully advantage of them to set up a relationship as best as possible to make a prediction of the remarking behaviors of the users. In this question, the most important thing is to select the proper model to make a relatively precise prediction.

2.3 About Question 3

While as for the third question, the difficulty lies in the final recommendation and selection, which means we have to move a step further to deeply analysis the users' interests. In this context, to cluster the vast books into different types and filter the less preferred ones becomes the most important question we should be concerned about.

三、 Assumptions

1） The books' ID is classified and coded as Dewey Decimal Number;

2） Readers give remarks of the books after they have read them through, which means as for the unread one, users cannot make remarks on them;

3） The reason why a user pays attention to other users is that they share the similar interests;

4） Tags of the internet only represent the times of classifications.

四、 Notation Explanations

a_i	Prevailing extent of the books
b	Type of the books
$c_i(i=1,2,3...9)$	Numbers of the each type of a target user;

五、 Model construction & Solution

5.1 Question 1

5.1.1 Calibration and the pre-screen of the data

Basically, we adopt the following four steps to calibrate and pre-filter the given statistics.

Step 1：To ensure that all the remarks and comments are made after the readers have indeed read the books, we filter out some remarks on the unread books to narrow down the potential influence on the later analysis;

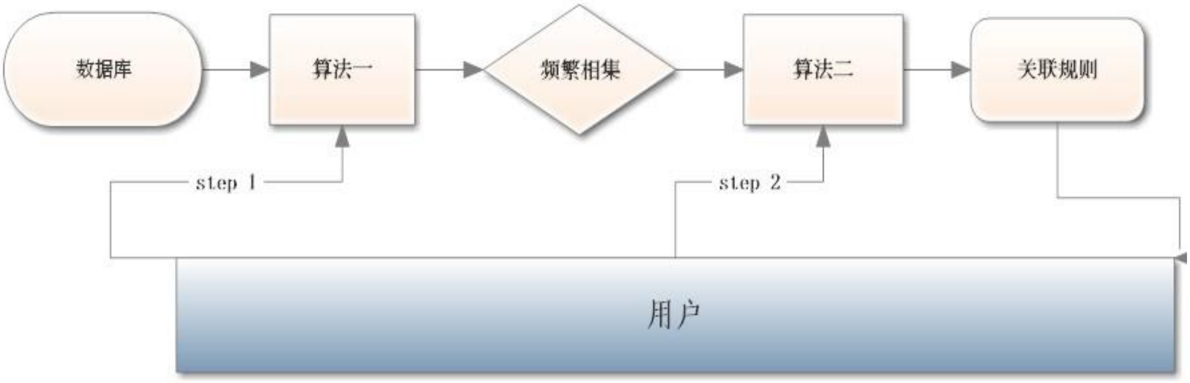
Step 2：We ignore the un-tagged books and take them as the newly published books to ensure that all the books are with tags;

Step 3：To complete the dimensions and form an integrated matrix, we make a missing value handling for the tags of the corresponding books;

Step 4：As for the repetitive book tags, we only select one of them to avoid the repetition effects.

5.1.2 Detail process of Factor One

Figure 5-1 The process of Heavy Data

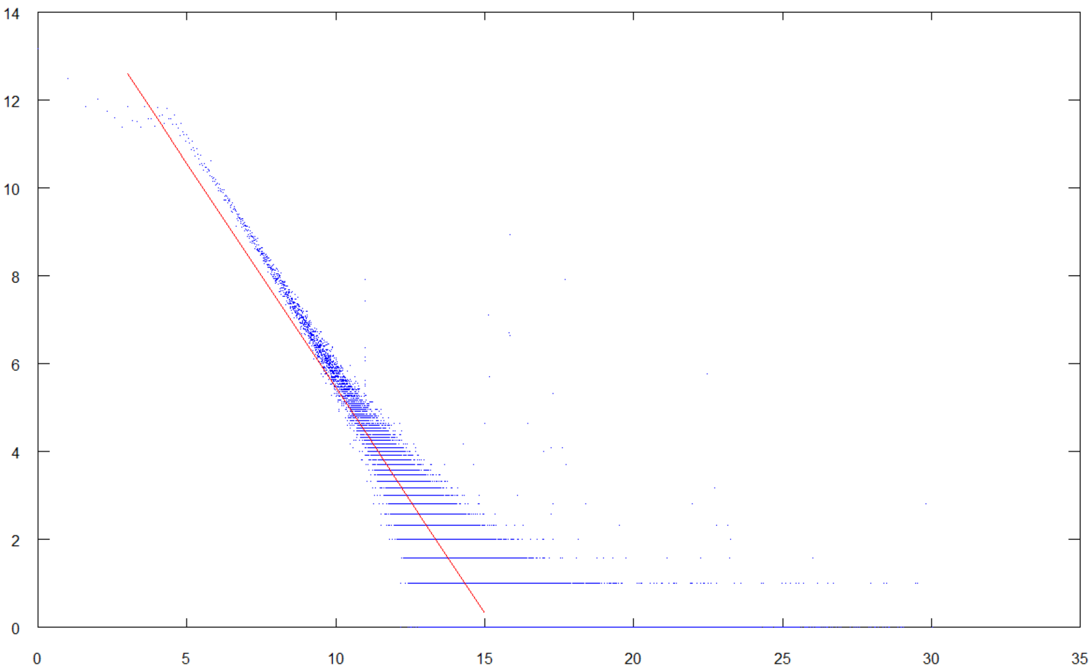


To simplify the selection process of the heavy statistics, we assume that the user activities and the prevailing extent of the books conform to the long-tailed distribution. In this context, when we define a tag which is used on a book by one user, then the corresponding prevailing extent of this book will add up by one level, which means:

$$a_i = a_i +$$

Therefore what we should do is to make an accumulated generating operation, and then we can obtain the prevailing extent of the corresponding books.

Figure 5-2 Long-tailed Distribution Sketch Map



To obtain a more direct impact on the users' remark, we make our efforts to cross analyze the remark history of the users' friends and find out that in the given statistics, only few friends of the users have commented and remarked on the related books. In this context, we excluded the influence on the users' remarks by their friends. Therefore, the prevailing extent plays much more important in the above remark process.

5.1.3 Detail process of Factor Two

According to the definition the Book ID, we find that the book tags of this net store are coded by Dewey Decimal Number, which means they are coded in the following chart:

Dewey Decimal Books Classification Method

Code	**Meaning**
000	**Pandect**
100	**Philosophy**
200	**Religion**
300	**Society & Science**
400	**Language**
500	**Natural Science & Mathematics**
600	**Technology (Applied Science)**
700	**Art**
800	**Literature**
900	**Geography History and other Supporting Disciplines**

Based on such characters, we can make a query about the reading history of the users and obtain the number of each type of the books to get the reading interest of the users, via the method of looping and superposition search. Taking the reading history of user 7245481 as an example, we can obtain the following chart:

Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8	Type 9
69	66	62	69	68	76	78	63	88

From the above chart, we can basically infer the basic reading hobbies of the user 724548, and it can be obtained that this user has a wide interest of the books and each type of the books conforms to a generally even distribution, however, the geography, history and the supporting discipline books are much more preferred by him/her.

In conclusion, we can get the second influence factor which may affect the remark behaviors of the users, which is the reading interest of the users.

5.1.4 Conclusion of Question 1

After the analysis and question 1, we concluded that there are two major factors which may affect the remark behaviors of the users,

1. One is the prevailing extent of the books remarked,

This factor can be explicated as the add-up of the book tags, and the loading times of each tag can be taken as the heat of this target book;

2. The other is the reading interest of the unread books.

This factor can be explained and represented by the types of books in the reading history of the users, and the larger the proportions each book takes up, the much more preference the readers tend to show.

5.2 Question 2

5.2.1Previous Preparation

As we adopt the five-score evaluation to evaluate the books, then, the scoring formula can be made as the following way:

Figure 5-2 Scoring Formula of the Coding Method

User ID	Books ID	Scores	Code of the Score
7245481	962729	4.0	（00010）
7625225	537793	3.0	（00100）
4891693	319726	5.0	（00001）
4891693	637116	2.0	（01000）
1388583	574530	1.0	（10000）

Add up the tags of the books, then we can obtain the heat of the target books, for example, the heat of Book ID **852102 can be represented as:**

Book ID	Heat1	Heat2	Heat3	Heat4	Heat5	Heat6	Heat7	Heat
852102	4770	2854	2069	3151	7539	6088	6957	33428

5.2.2 Model Construction

BP（Back Propagation）Neural Network is proposed by a scientists group led by Rumelhart and McCelland in 1986, which is a multilayer feed forward network trained by Error Back-Propagation algorithm and is one of the most widely-used Neural Networks. BP network can learn and store a lot of input - output model mapping relations, without prior revealing the description of the mathematical equations of the mapping relationships. And its learning rule is to use the steepest descent method to constantly adjust the network weights and threshold by back propagation to minimize the error sum of squares of the network. The topological structure of the BP Neural network is consisted of input layer, hidden layer and the output layer.

Figure 5-3 Diffusion Map of BP Neural Network

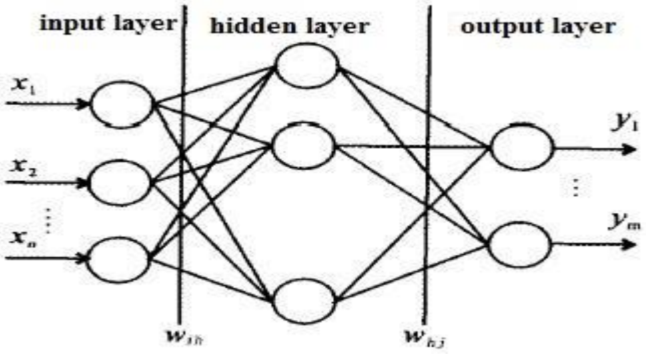
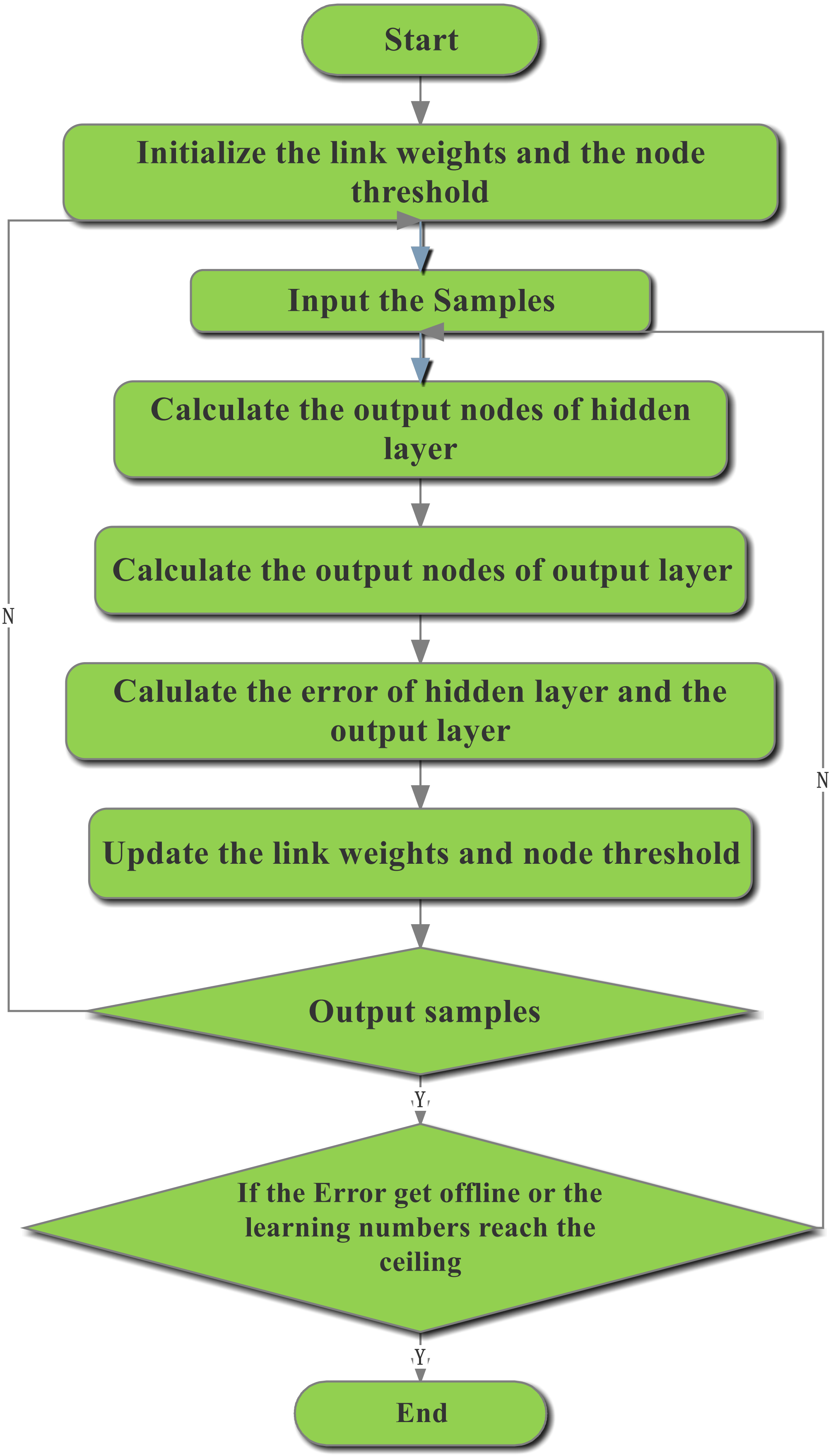


Figure 5-4 Operation Flow chart of BP Neural Network


Start
Initialize the link weights and the node threshold
Input the Samples
Calculate the output nodes of hidden layer
Calculate the output nodes of output layer
Calulate the error of hidden layer and the output layer
Update the link weights and node threshold
Output samples
N
Y
If the Error get offline or the learning numbers reach the ceiling
N
Y
End

In this paper, we take the book type b , book heat a_i , and the numbers of the books that the users read in their reading history $c_1, c_2 ... c_9$ as the input elements, by the process of normalization used the following formula:

$$y = \frac{x - MinVaule}{Maxvaule - MinVaule}$$

We obtained the following 10 nerve cells.

After that, we adopt the AHP method and select 6000 samples from the books which have been remarked as the training target, and the detailed initial parameters of the BP Neutral Network are shown in the following chart:

Initial Parameters	Detailed Values
Iteration times	100
Learning Rate	0.1
Target Accuracy	0.0004
Initial weights	randomly Given

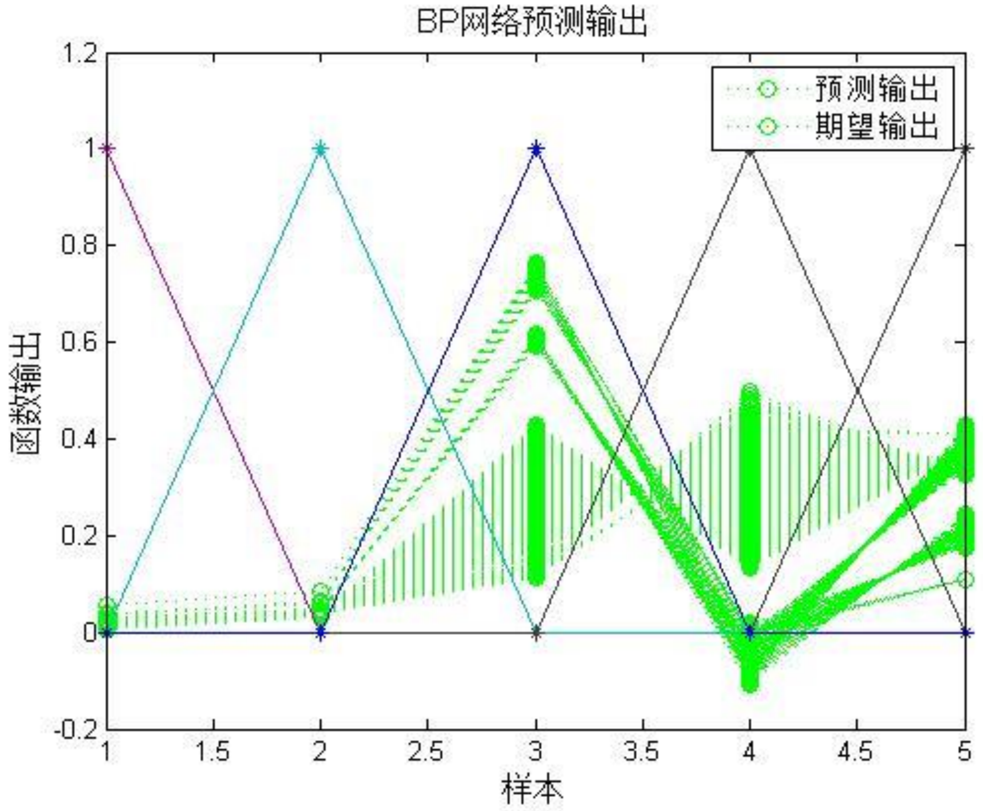
We then carry out the training process based on the above initial parameters and input Nerve Cells.

5.2.3 Solution of the Model

Using the well trained Network to make prediction to the books, we obtained the following result:

User ID	Book ID	Predicted Score	User ID	Book ID	Predicted Score
7245481	794171	4. 0	4156658	134003	4. 0
7245481	381060	4. 0	**4156658**	**443948**	4. 0
7245481	776002	4. 0	**5997834**	**346935**	3. 0
7245481	980705	4. 0	**5997834**	**144718**	4. 0
7245481	354292	4. 0	**5997834**	**827305**	4. 0
7245481	738735	4. 0	**5997834**	**219560**	4. 0
7625225	473690	3. 0	**5997834**	**242057**	4. 0
7625225	929118	3. 0	**5997834**	**803508**	3. 0
7625225	235338	3. 0	**9214078**	**310411**	4. 0
7625225	424691	3. 0	**9214078**	**727635**	5. 0
7625225	916469	3. 0	**9214078**	**724917**	4. 0
7625225	793936	3. 0	**9214078**	**325721**	4. 0
4156658	175031	5. 0	**9214078**	**105962**	3. 0
4156658	422711	4. 0	**9214078**	**235338**	5. 0
4156658	585783	5. 0	**2515537**	**900197**	4. 0
4156658	412990	5. 0	**2515537**	**680158**	2. 0
			2515537	**770309**	4. 0

Figure 5-5 Output of the BP Prediction and Expectation



In the meantime, which can be shown in the above picture, we output the error of prediction in the Neural Network, and concluded that with the increase of the number of the samples, the output value becomes more accordant with the fact. And the following BP Neural Network Figure can also prove this conclusion.

Figure 5-6 Error Figure of the Prediction by BP Neural Network

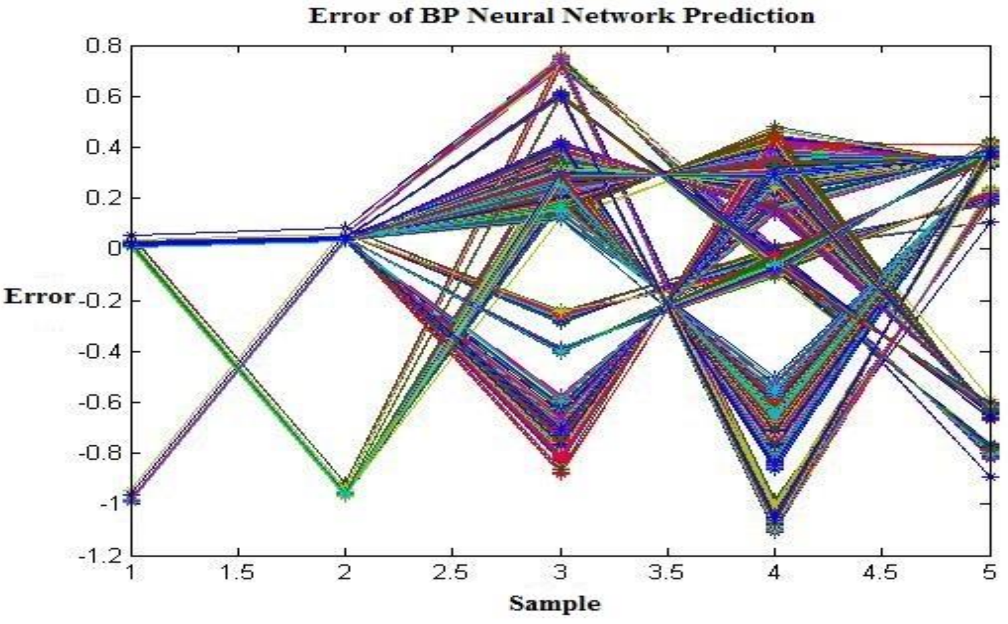
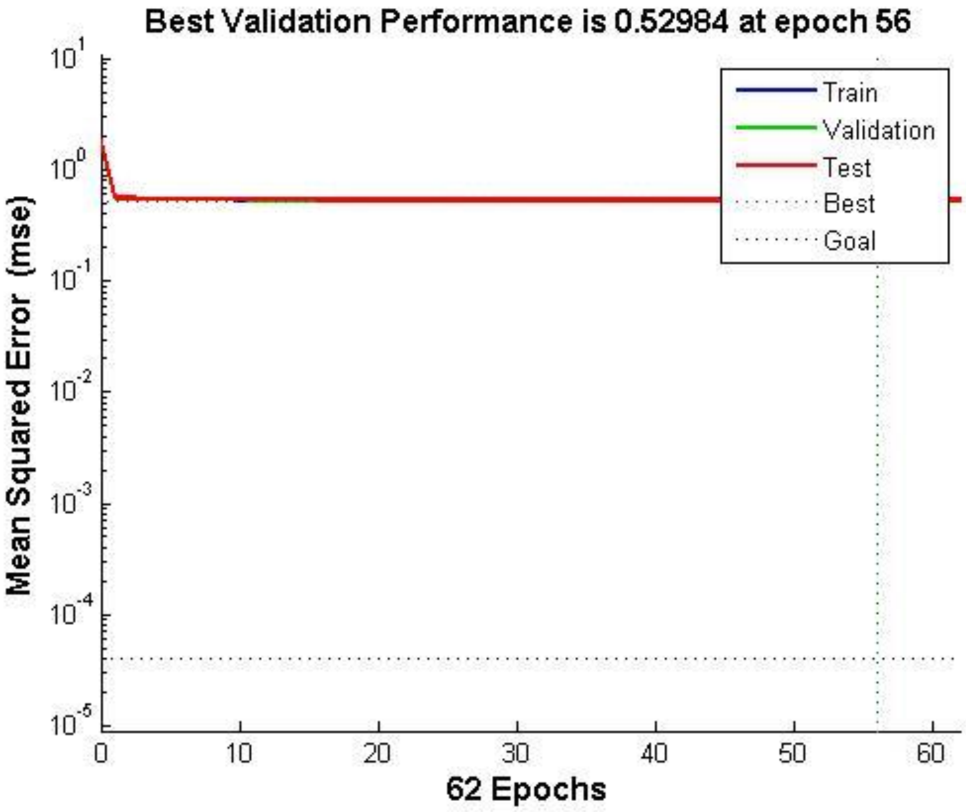


Figure 5-7 Convergence Pic. Of the BP Neural Network Prediction



5.3 Question 3

5.3.1 Clustering based on the user' s interest

Input : the users' interest to each type of the books c_i , the matrix of the concerned friends S, threshold of the difference of the set d , and the total number of the item classification G;

Output : The number of the users based on their different interests- Cluster.

Supposing that the total number of the items in the category set N is n, then we can calculate the value of I until all the values of the "friends "has been used, in the end, we can calculate the total number of the each group and realign the categories marked as N^i

Next, we set up an initial item set C_i for each item of the set, and each item set has only one item;

Then, we can use the following equation to calculate the convergence difference of the different item sets, which can be expressed as

$$SFD(C_1,C_2)=\frac{K-I}{K\times G}$$

If the convergence difference of different items SFD is less than the value of threshold d， then we can combine the two item sets as a whole one。

And this final combined item set is the Cluster， which includes m item class， and each set has its own number to remove the isolated item.

5.3.2 Collaborative Fitering

According to the user clustering analysis, we can compute the relative error consistent with the recommended target level of interest, and the mathematical equation is:

$$MSE=abs(\frac{c_1^1}{\sum_{i=1}^{9}c_i^1}-\frac{c_1^2}{\sum_{i=1}^{9}c_i^2})$$

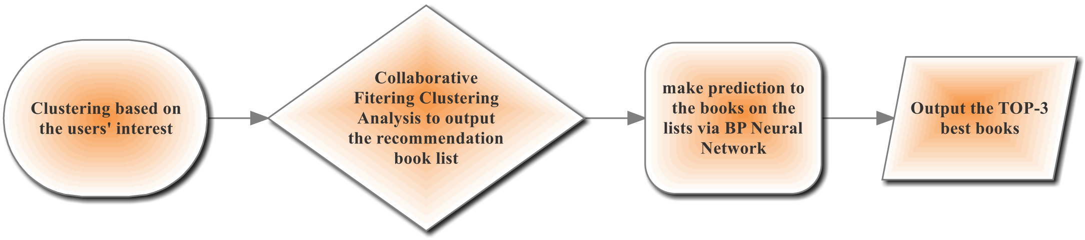
Add up the relative errors of each group, we then get the interest accumulated degree of difference, and the relatively smaller difference can be taken as the nearest neighbors. After that, we can recommend a more than 3 books TOP book lists to the users based upon their nearest neighbors ‘statistics.

And then, we adopt the Model 2 to make a prediction score to the recommendation books and form a ranking descending ranking list, the TOP 3 books in this book list can be seen as the TOP 3 best books to be recommended to the target users.

5.3.3 Solution

On the basis of Clustering as well as the collaborative filtering analysis, the given statistics can be narrowed to a small enough range. In this context, we can make a prediction on the users' remarks on the unread books by BP Neural Network and recommend TOP 3 books for the users. The detailed process can be seen in the following flow chart,

Figure 5-8 Analysis of the Clustering method based on the collaborative filtering


Clustering based on the users' interest
Collaborative Fitering Clustering Analysis to output the recommendation book list
make prediction to the books on the lists via BP Neural Network
Output the TOP-3 best books

Via the above flow chart we can obtain the following TOP-3 recommendation book lists:

User ID	7245481	7625225	4156658	5997834	9214078	2515537
Recommendation Book ID 1	908608	801049	730901	723581	424691	424691
Recommendation Book ID 2	922764	803508	736512	776002	418051	485254
Recommendation Book ID 3	936585	832057	745929	770309	432074	484434

六、Evaluation of the Model

6.1 Advantage

1. The Model proposed in this paper can be used to deal with massive complicated statistics to find out the possible rules theses data may follow and then make a classification, which means we can use this model to deeply explore the inner relations of the provided statistics for the following study;

2. BP Neural Network enjoys a strong nonlinear mapping ability, self-learning and adaptive ability, and generalization and fault tolerance ability, which can make an quick and rather precise evaluation to the target users;

3. Clustering based on the collaborative filtering method can effectively reduce the dimensions of big data, and accelerate the process of searching for the nearest neighbors, hence making the results enjoy a strong regularity.

6.2 Disadvantage

1. It's not intelligent enough for the Big Data Mining method;

2. The convergence rate of BP Neural Network is a little slow and dependent on the number and the precision of the samples;

3. The center of clustering method is difficult to determine and control according to their own requirements.

七、Extension and the Improvement of the Model

7.1 Improvement of the Model 模型的改进

For the BP Model, it can be improved by RBF Neural Network in theory, and it may has a better effect; while as for the BP Neural Network itself, it can be implemented by Genetic Algorithm to optimize the process of hidden layer to be more precise.

As for the clustering method, SOM Neural Network is also a good choice, and the use of topologic structure to explicit the kinds and numbers of the clustering would be preferred to further improve the model itself.

7.2 Extension of the Model 模型的推广

The Model we proposed in this paper can not only be used in the books remark and prediction system, but also goes for the series of plays, movies and games and other recommendation system.

Additionally, only by a little improvement, this model can be used in the prediction field, for example, the stock market prediction and the scoring of safety factors and so on.

八、Reference

【1】 Shoukui Si，Add ink algorithm and applied mathematics（Edition 1），Peking：National defence of Industry Press，2011.08

【2】 Feng Shi et al，30 Cases Analysis of matlab Neural Network，Bejing：Beijing University of Aeronautics and Astronautics Press，2010.4

【3】Liang Xiang，Practice of Recommendation System，Beijing，people's posts and telecommunications publishing Press

【4】Zhong Yao, Jia wei, Yue Wu，Collaborative filtering recommendation algorithm based on high-dimensional sparse data clustering，Beijing University of Aeronautics and Astronautics Press，2008

【5】Tao Peng，Analysis and Evaluation of the Online Book Lists，2012.14

九、Appendix

```
load(’outscore.mat’)
a=outscore;
a=sum(outscore,2);
for i=1:189791
    switch(a(i))
        case 1
            outscore(i,:)=[1 0 0 0 0];
        case 2
            outscore(i,:)=[0 1 0 0 0];
        case 3
            outscore(i,:)=[0 0 1 0 0];
         case 4
            outscore(i,:)=[0 0 0 1 0];
         case 5
            outscore(i,:)=[0 0 0 0 1];
    end
end
```

```
% 双隐含层 BP 神经网络
%% 清空环境变量
clc
clear

%% 训练数据预测数据提取及归一化
%下载输入输出数据
load(’llscore.mat’)
load(’outscore.mat’)
load(’predict.mat’)
load(’newscore.mat’)
load(’newoutput.mat’)
input=newscore;
output=newoutput;

%找出训练数据和预测数据
input_train=input’;
input_test=llscore(60001:66791,:)’;
input_predict=predict’;
output_train=output’;
output_test=outscore(60001:66791,:)’;
```

```
%选连样本输入输出数据归一化
[inputn,inputps]=mapminmax(input_train);
[outputn,outputps]=mapminmax(output_train);

%% BP网络训练
% %初始化网络结构
net=newff(inputn,outputn,[5 5]);

net.trainParam.epochs=1000;
net.trainParam.lr=0.1;
net.trainParam.goal=0.004;

%网络训练
net=train(net,inputn,outputn);

%% BP网络预测
%预测数据归一化
inputn_test=mapminmax('apply',input_test,inputps);

%网络预测输出
an=sim(net,inputn_test);

%网络输出反归一化
BPoutput=mapminmax('reverse',an,outputps);

%% 结果分析

figure(1)
plot(BPoutput,':og')
hold on
plot(output_test,'-*');
legend('预测输出','期望输出')
title('BP网络预测输出','fontsize',12)
ylabel('函数输出','fontsize',12)
xlabel('样本','fontsize',12)
%预测误差
error=BPoutput-output_test;


figure(2)
plot(error,'-*')
title('BP网络预测误差','fontsize',12)
ylabel('误差','fontsize',12)
xlabel('样本','fontsize',12)
```

```
figure(3)
plot((output_test-BPoutput)./BPoutput,'-*');
title('神经网络预测误差百分比')

errorsum=sum(abs(error));
%预测
inputn_predict=mapminmax('apply',input_predict,inputps);
an=sim(net,inputn_predict);
predict_simu=mapminmax('reverse',an,outputps);



%% 清空环境变量
clc
clear

%% 训练数据预测数据提取及归一化
%下载输入输出数据
load data input output

%从 1 到 2000 间随机排序
k=rand(1,2000);
[m,n]=sort(k);

%找出训练数据和预测数据
input_train=input(n(1:1900),:)';
output_train=output(n(1:1900));
input_test=input(n(1901:2000),:)';
output_test=output(n(1901:2000));

%选连样本输入输出数据归一化
[inputn,inputps]=mapminmax(input_train);
[outputn,outputps]=mapminmax(output_train);

%% BP 网络训练
% %初始化网络结构
net=newff(inputn,outputn,5);

net.trainParam.epochs=100;
net.trainParam.lr=0.1;
net.trainParam.goal=0.00004;

%网络训练
```

```
net=train(net,inputn,outputn);

%% BP网络预测
%预测数据归一化
inputn_test=mapminmax('apply',input_test,inputps);

%网络预测输出
an=sim(net,inputn_test);

%网络输出反归一化
BPoutput=mapminmax('reverse',an,outputps);

%% 结果分析

figure(1)
plot(BPoutput,':og')
hold on
plot(output_test,'-*');
legend('预测输出','期望输出')
title('BP网络预测输出','fontsize',12)
ylabel('函数输出','fontsize',12)
xlabel('样本','fontsize',12)
%预测误差
error=BPoutput-output_test;


figure(2)
plot(error,'-*')
title('BP网络预测误差','fontsize',12)
ylabel('误差','fontsize',12)
xlabel('样本','fontsize',12)

figure(3)
plot((output_test-BPoutput)./BPoutput,'-*');
title('神经网络预测误差百分比')

errorsum=sum(abs(error))
```