MathorCup 全球大学生数学建模挑战赛

承　诺　书

我们仔细阅读了 MathorCup 全球大学生数学建模挑战赛的规则.

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的, 如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们参赛选择的题号是（从 A/B/C 中选择一项填写）：　B　

我们同意组委会可以公开发布论文到校苑数模网：　是　（是/否）

我们的参赛报名队号：　10352　

参赛队员 ：1.　邢云飞　

2.　张丽娜　

3.　宋迎召　

指导教师或指导教师组负责人 ：　廖川荣　

日期：　2014　年 5　月 28 日

评委一评分，签名及备注	队号： 10352	评委三评分，签名及备注
评委二评分，签名及备注	选题： B	评委四评分，签名及备注

题目：**推荐书籍**

摘要

随着信息技术和互联网技术的发展，信息逐渐由匮乏时代转入过载时代。图书市场也是这样的步伐，对于阅读者如何从海量图书中选到自己喜欢并且高质量的图书是件困难的事；同时对于作者来说，如何使自己的书脱颖而出也是件非常困难的事情。本文根据所提供数据，深度挖掘数据之间的联系性，建立出行之有效的模型来预测评分和推荐书籍。

针对问题一，主要是进行大数据的信息挖掘，本文通过关联规则应用于高维，海量的数据探寻中，通过降低维度，查询资料发现书籍标签热度服从长尾分布，建立数据间的映射表，通过缺失值处理方法补充成完整矩阵的方法，找出与用户评分有关的强关联影响想因素。通过本文综合运用大量的大数据挖掘技术，最终确定了影响用户对书籍的评分影响因素有：1.用户的阅读兴趣 2.书籍的流行度

针对问题二，在解决了问题一的基础上，并且认为文中给出的图书 ID 是按照杜威十进制数来编码，通过抽样算法抽取出 60000 个用户读书类型的样本数据，以及对应评价书籍的热度作为输入端，选取了 6000 个已评分的书籍记录作为神经网络的校验。此时我们将评分与影响因素的关系看成一个黑盒子，使用 BP 神经网络对输入输出数据进行训练，最好用训练好的网络进行评分预测。完美的发挥了 BP 神经网络的非线性系统的优越性，较为准确的预测出用户的评分。（具体评分见文章）

针对于问题三，运用基于聚类的协同过滤方法，以用户的兴趣爱好为聚类中心，将有相同爱好的读者聚集在一起。之后，再找出与推荐对象的最近邻居，根据最近邻居来协同帮助发现用户的隐性信息，从而选择出 TOP-3，将这三本书推荐给用户阅读。（具体推荐书籍见文章）

关键词：大数据挖掘　长尾分布　BP 神经网络　聚类　协同过滤分析

一、 问题重述

1.1 问题的背景

随着网络的普及，图书出版业也迎来了爆棚时代，读者面临的信息量越来越大，可供选择的书籍也越来越多，此时如何选到一本心满意足的书籍已经变得不那容易。应于时代的要求，个性化推荐应运而生，它从用户的历史数据和用户的社交行为数据中发现用户的 “兴趣”，采取推荐的方式将信息呈现在用户面前，使用户尽量快的从海量的信息中找到自己感兴趣的书籍。然而，目前国内外对于图书评价的研究，无论在理论上还是实际中都相对落后。目前，对于图书评价和图书的推荐仍然处于定性的分析层面上。所以，有必要通过用户的资料以及历史行为对书籍评分进行预测并且实现较为准确的书籍推荐系统。

1.2 问题的提出

根据题目给出的数据以及要求，本体可以归纳为以下三个问题：

1. 挖掘题目中的数据内在联系。并且观察评分与数据间的关系。从中分析出对于用户评分的影响因素‘
2. 根据问题一的影响因素，建立适当的预测模型对表中用户未评过分的书籍进行评分。
3. 利用用户的社交数据，使用协同过滤的方法给用户推荐符合兴趣爱好的书籍。

二、 问题的分析

2.1 问题一的分析

题干中明确的说明从“数据”中挖掘信息解决以下的问题。对于问题一，就必须从海量的数据中去发现联系。所以问题的关键突破点在于如何从几十万的数据中找到影响评分的因素。所以，问题一的主要任务便是进行大数据挖掘，从中找出影响用户评分的因素。

2.2 问题二的分析

问题二的基础便是问题一中所分析出来的影响因素，问题二要解决的就是如何利用这些影响评分因素的大量数据，通过数据之间简历一种怎样的关系去对用户评分行为进行预测。所以问题二要选定合适的处理模型来进行评分预测。

2.3 问题三的分析

问题三，在问题一的分析基础上要更进一步的根据用户的兴趣，在大量的书籍中选出用户最感兴趣的书籍进行推荐是难点。将大量的书籍聚类并借助歪理来过滤那些用户不喜欢的书籍是这一问的重点。

三、 模型的假设

1） 书籍的 ID 是按照杜威十进分类法来进行分类的
2） 用户必须在看过书籍后才可以对书籍评分
3） 用户关注的好友大多数是因为兴趣相近
4） 网站的标签数仅代表被归类的次数

四、 符号说明

a_i	书籍的流行度
b	书籍的类型
$c_i(i=1,2,3...9)$	用户所读每种类型书的本数
MSE	相对误差

五、 模型建立与求解

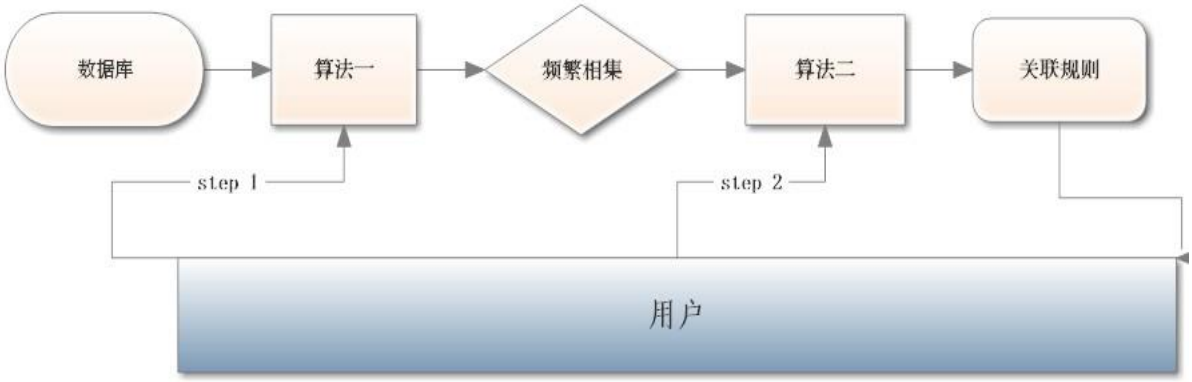
5.1 问题一的求解

5.1.1 数据的校验和前期筛选

Step 1：确认所有评分用户在看过书籍之后再评分，对于有些用户未看过该书籍就评分的作为错误数据，为避免对后期分析影响故将其删除
Step 2：确认所有书籍都有标签，没有标签的书籍当作刚发布，没有任何历史记录我们不予以考虑
Step 3：将书籍对应的标签数据作缺失值处理，补全维数形成一个完整的矩阵
Step 4：将重复的书籍标签或者阅读历史取其中一个

5.1.2 因素一的处理和归纳

图 5-1 大数据挖掘流程



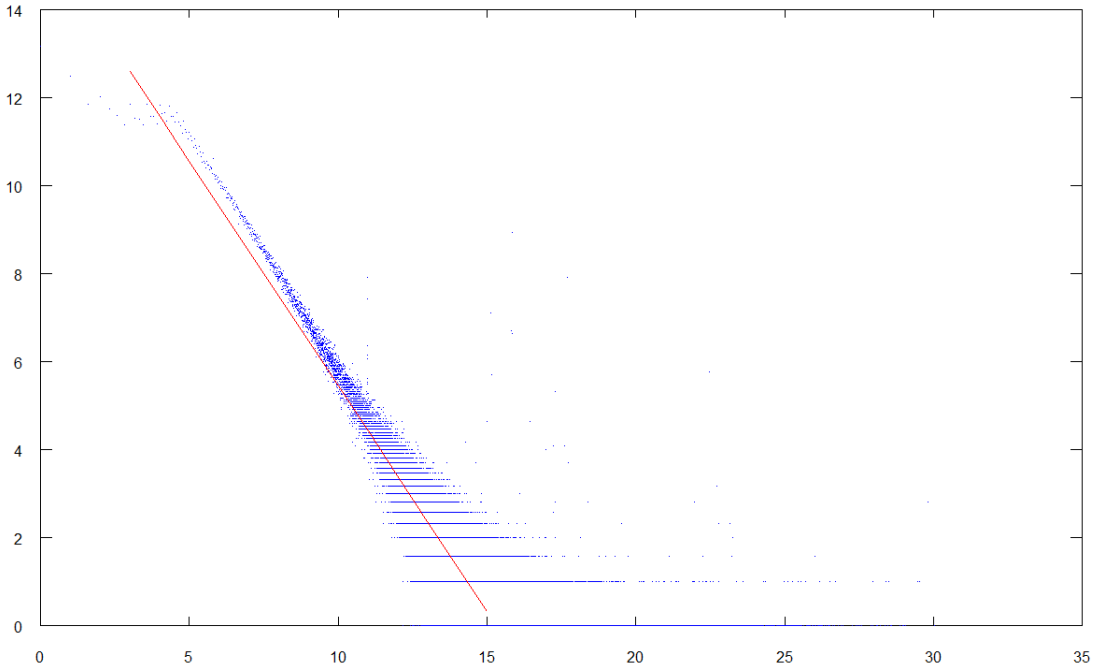
通过算法使得数据产生适应性，对大批量数据进行简化筛选过程。通过查询资料知道在用户集中用户活跃度和书籍流行度都符合长尾分布。因此我们定义一个标签被读者用在一本书籍上，那么这本书籍的活跃度就被赋值为原本活跃度加一。

即为：

$$a_i = a_i + 1$$

通过对书籍标签数作一次累加得出的数可作为书籍的流行度。

图 5-2 长尾分布图示意



我们试图从用户好友与用户的评分历史中找出交集，使得用户评分影响因素来得更直接。可我们发现，带题中给出要预测用户的好友之中几乎没有好友对书籍

评过分，由此我们可以排除好友打分对于用户评分的影响。因此在以上部分中可以发现书籍的流行度对于评分有着较为重要的影响

5.1.3 因素二的处理和归纳

通过查找书籍 ID 的含义，我们确定该网上书店采用的书籍标签是按照杜威十进图书分类法去确定 ID。

杜威十进图书分类法
000　总论
100　哲学
200　宗教
300　社会科学
400　语言
500　自然科学和数学
600　技术（应用科学）
700　艺术、美术和装饰艺术
800　文学
900　地理、历史及辅助学科

我们根据此类特点可以查询用户的阅读历史，通过对用户阅读历史的循环搜索和叠加，我们可以得到用户阅读每类书籍的本书，从而可以确定用户的阅读兴趣和爱好。比如用户 7245481 的阅读历史

表 5-1 阅读兴趣

第一类	第二类	第三类	第四类	第五类	第六类	第七类	第八类	第九类
69	66	62	69	68	76	78	63	88

可以从这份表格中大体推断出用户 7245481 的阅读喜好，该用户爱好广泛，各类书籍读的都比较均匀，但地理、历史及辅助学科较为突出一些，可得出该用户偏爱第九类书籍。

综合以上的叙述，可以得出影响用户评分的第二个因素，也是最重要的影响因素即为：用户的阅读兴趣。

5.1.4 问题一的归纳总结

对于用户评分的影响因素有

1.被评价书籍的流行度

在数据中体现为书籍标签的加载总和，书籍标签的加载次数即为该书籍的热度。

2.用户的阅读兴趣

采用户阅读历史中书籍类型的分布来表示，阅读过的书籍中某类型比例越大则代表用户对于那类书籍比较偏爱。

5.2 问题二的求解

5.2.1 模型的前期准备

由于评分制度是采用五分制制度，所以制定如下编码的计分方式：

表 5-2 编码计分方法

用户 ID	评价书籍 ID	评价得分	编码得分记为
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）

将图书标签数累加当作该书籍热度：

如表 5-3 书籍 852102 热度为

书籍 ID	标签 1	标签 2	标签 3	标签 4	标签 5	标签 6	标签 7	热度为
852102	4770	2854	2069	3151	7539	6088	6957	33428

5.2.2 模型的建立

BP（Back Propagation）网络是 1986 年由 Rumelhart 和 McCelland 为首的科学家小组提出，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。BP 网络能学习和存贮大量的输入-输出模式映射关系，而无需事前揭示描述这种映射关系的数学方程。它的学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP 神经网络模型拓扑结构包括输入层（input)、隐层(hidden layer)和输出层(output layer)

图 5-3 BP 神经网络传播图

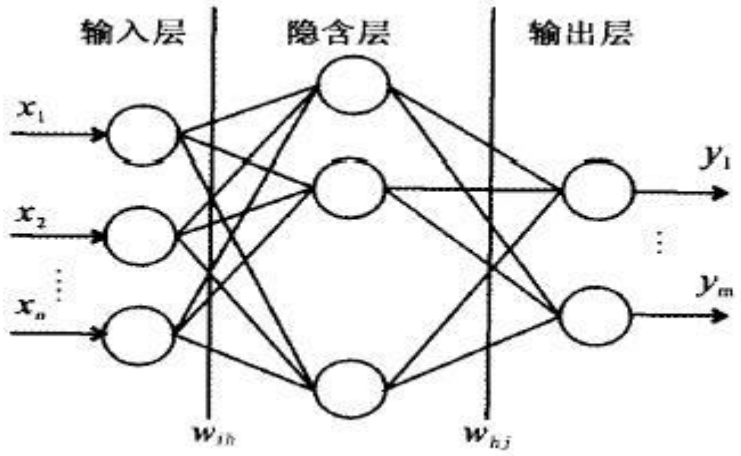
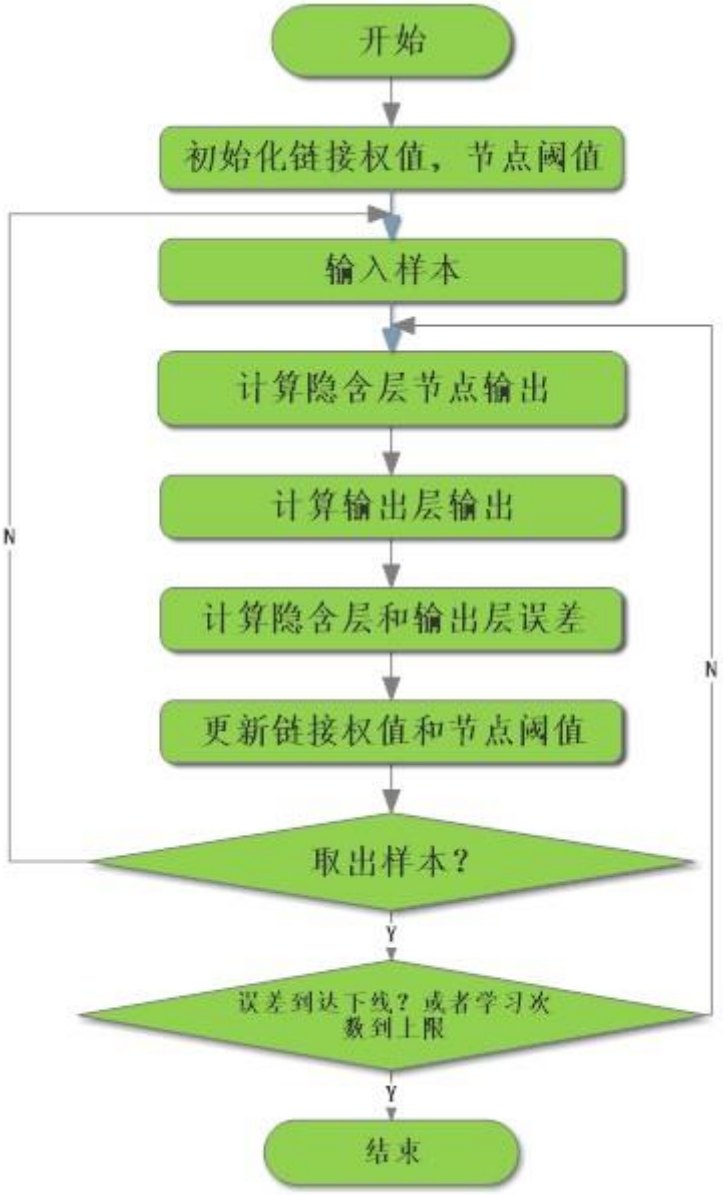


图 5-4 BP 神经网络运行流程图



文中将书的类型b，书籍的热度a_i，以及用户的阅读历史中各类书的本数$c_1,c_2...c_9$作为输入元素，通过归一化函数：

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

得到了 10 个输入神经元，通过层次取样法从已经评过分的书籍中选取 6000 个样本作为训练目标。设置 BP 神经网络的各初始参数为：
迭代次数 100 次；
学习速率 0.1；
学习的目标精度 0.0004；
初始权值随机给定；

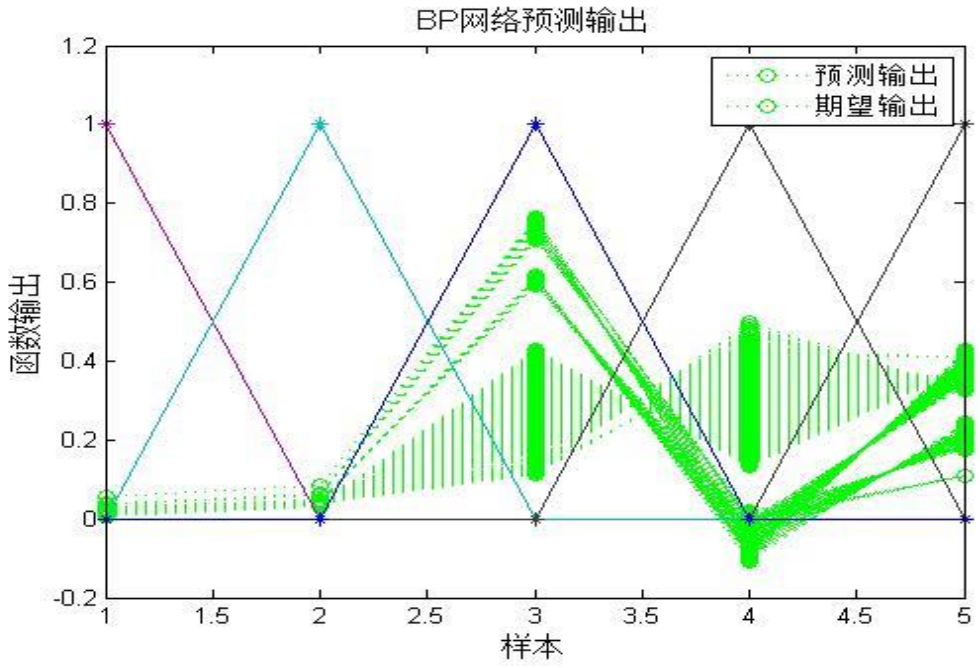
通过以上的输入神经元以及初始值设置，对神经网络进行训练。

5.2.3 模型的求解

使用训练好的网络对用户评分的预测为：

用户 ID	书籍 ID	预测评分
7245481	794171	4.0
7245481	381060	4.0
7245481	776002	4.0
7245481	980705	4.0
7245481	354292	4.0
7245481	738735	4.0
7625225	473690	3.0
7625225	929118	3.0
7625225	235338	3.0
7625225	424691	3.0
7625225	916469	3.0
7625225	793936	3.0
4156658	175031	5.0
4156658	422711	4.0
4156658	585783	5.0
4156658	412990	5.0
4156658	134003	4.0
4156658	443948	4.0
5997834	346935	3.0
5997834	144718	4.0
5997834	827305	4.0
5997834	219560	4.0
5997834	242057	4.0
5997834	803508	3.0
9214078	310411	4.0
9214078	727635	5.0
9214078	724917	4.0
9214078	325721	4.0
9214078	105962	3.0
9214078	235338	5.0
2515537	900197	4.0
2515537	680158	2.0
2515537	770309	4.0
2515537	424691	3.0
2515537	573732	3.0
2515537	210973	3.0

图 5-5BP 预测输出和期望输出



同时在神经网络中输出 BP 神经网络的预测误差：可见随着样本数的增多，BP 神经网络的输出与实际越符合。结合下图的 BP 神经网络的预测误差图也可以说明该问题。

图 5-6 BP 神经预测误差

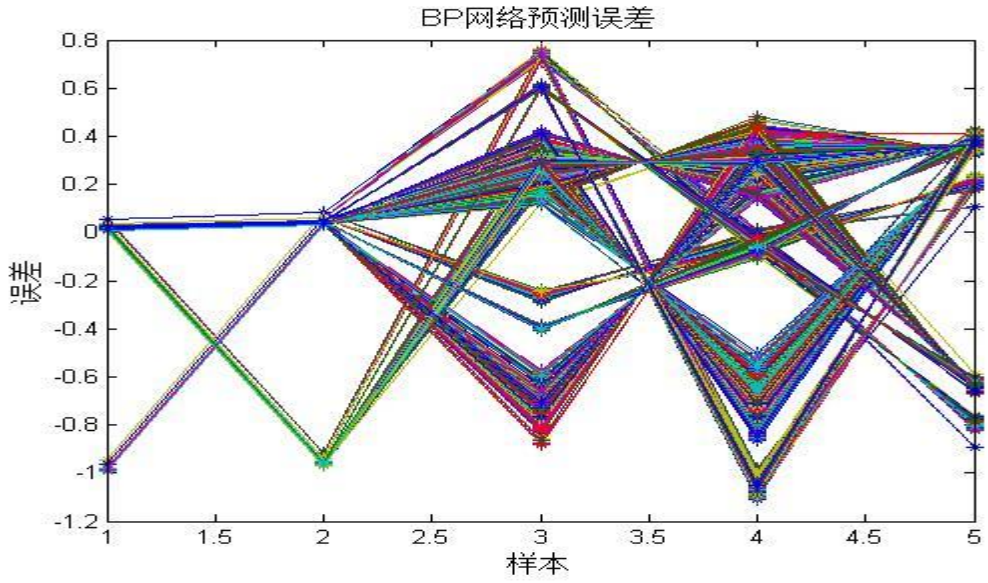
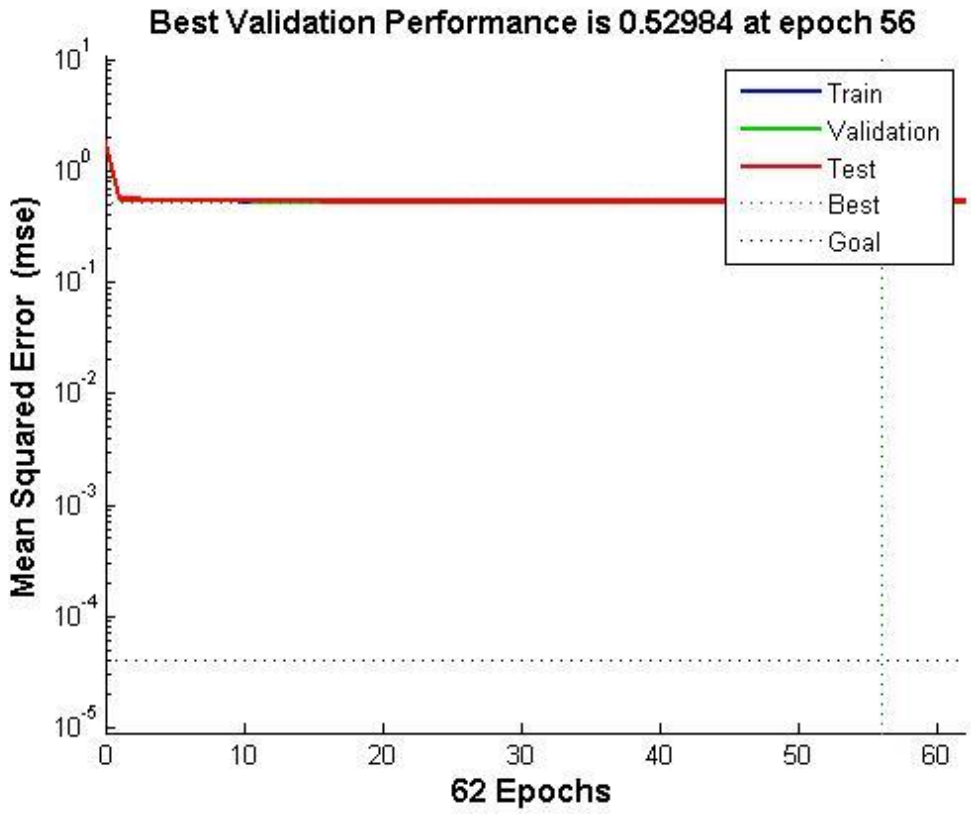


图 5-7 神经网络预测收敛图



5.3 问题三的求解

5.3.1 基于用户兴趣的聚类

输入：用户对于每一类书籍的兴趣度 C_i，关注好友的矩阵 S，集合度差异度阈值 d，项目类别总数 G。

输出：基于兴趣的人群 Cluster.

设类别集 N 中项目的总数为 n，依次计算项目中 I 直至所有好友取尽，得到每组的人数，将类别重新排列记为 N^i

为项目集中的每一个项目创建一个初始的项目集合 C_i，每个项目集合中只有一个项目

计算不同的项目集合的差异度，计算公式为：

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

计算不同项目集之间的差异度，如果差异度 SFD 小于阈值 d，则将两个项目集合合并。

通过项目集合的合并后得到的项目集合构成项目簇 Cluster，项目簇中包括 m 个项目类，每个项目中记录了属于该项目的编号，将孤立的类除去。

5.3.2 协同过滤

根据聚类中的用户进行分析，计算聚类中的用户与推荐对象的兴趣匹配程度相对误差，数学表达式为：

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

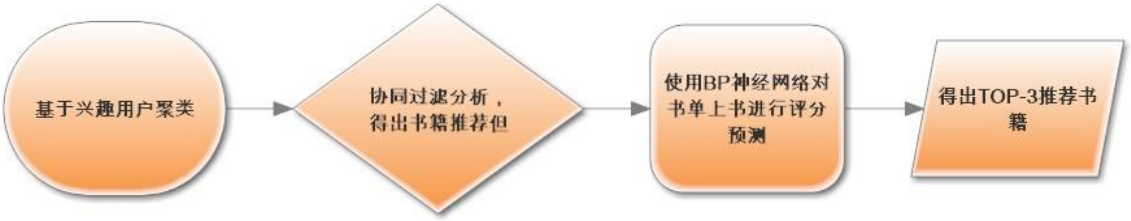
将每一组的相对误差累加得出兴趣差异度，差异度小的当作最近邻居，通过最近邻居的阅读历史产生对用户推荐书单由于此时的对每个用户推荐书单将远大于 3 本。此处使用模型二，对产生的推荐书单进行一个预测评分，将书单按照预测评分进行排列。选取预测评分较高的三本书作为 Top-3 推荐给特定的用户。

5.3.3 问题三的求解

在聚类的基础上，加上协同过滤分析，数据得以筛选到一个很小的范围内。通过 BP 神经网络的预测得到用户的预估评分，再选 TOP-3 来推荐给用户.

基于聚类的协同过滤方法流程如下图所示：

图 5-8 基于聚类的协同过滤分析



得出以下推荐结果

用户 ID	7245481	7625225	4156658	5997834	9214078	2515537
推荐书籍 1	908608	801049	730901	724396	418051	424691
推荐书籍 2	922764	803508	736512	778269	436304	485254
推荐书籍 3	936585	832057	745929	885390	487373	484434

六、模型的评价

6.1 模型的优点

1.本文的模型可以处理大量的复杂数据，可以在复杂数据中找到规律并加以归类。能够利用模型挖掘大数据的深层信息，并通过深层信息找到一些隐性的内在联系。

2. BP 神经网络有着较强的非线性映射能力、自学习和自适应能力、泛化能力和容错能力，能够快速较为准确的训练并且预测出用户评分

3.基于聚类的协同过滤方法能够有效的缩小大数据的维度，加速后期寻找最近邻居的过程，并且使得结果看起来规律性较强。

6.2 模型的缺点

1.大数据挖掘技术不够智能化

2.BP 神经网络收敛速度较慢而且比较依赖于样本的量和准确性

3.聚类方法的聚类中心点难以确定，很难控制按自己的要求聚类

七、模型的改进和推广

7.1 模型的改进

对于 BP 模型，现在理论研究可以改进为 RBF 神经网络来替代效果更好，BP 神经网络的本身也可以通过遗传算法来优化隐含层的处理过程，从而获得更高的精度。

聚类的方法也可以通过 SOM 神经网络来实现，通过拓扑结构来表明聚类的种类以及个数更加明显

7.2 模型的推广

该模型不仅仅可以用于图书的评分预测和推荐系统，同时还可以应用于电视剧、电影、游戏等一些推荐系统中

该模型稍加改造便可以应用在一些预测地方，比如股票市场的预测和安全系数评分等

八、参考文献

【1】司守奎，数学加墨算法与应用（第 1 版），北京：国防工业出版社，2011.08
【2】史峰等，matlab 神经网络 30 个案例分析，北京：航空航天大学出版社，2010.4
【3】项亮，推荐系统实践，北京，人民邮电出版社
【4】姚忠、魏佳、吴跃，基于高维稀疏数据聚类协同过滤推荐算法，北京航空航天大学，2008
【5】彭陶，网络图书排行榜评价指标探析，2012.14

九、附录

```
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))
```