Table of Contents

[1. Background 2](#_Toc459386542)

[2. Design 3](#_Toc459386543)

[3. Data Description 6](#_Toc459386544)

[4 Prediction migration rates 7](#_Toc459386545)

[4.1 Migration Data 8](#_Toc459386546)

[4.2 Selecting features 9](#_Toc459386547)

[4.3 Building the model: 15](#_Toc459386548)

[4.3.1 Bagging 15](#_Toc459386549)

[4.3.2 Linear Regression 18](#_Toc459386550)

[4.3.3 Random Forest 18](#_Toc459386551)

[4.4 Future Prediction 22](#_Toc459386552)

[5. Neural Network for individual prediction: 24](#_Toc459386553)

[5.1 Methods: 24](#_Toc459386554)

[5.2 Data Preparation: 25](#_Toc459386555)

[5.3 PCA 26](#_Toc459386556)

[Reference 29](#_Toc459386557)

# 1. Background

With the fast growing export industry, 140 million of Chinese citizens are under rural-to-urban migrants (10% of the total population) in leaving their home provinces for seeking for better opportunities for jobs and better environment for families(1). According to the International Labor Organization. Countless reasons such like economic situation in difference provinces, cultural differences, accents, distances from home and policies can cause them for migrations. To analyze this massive migration, a method that can select relevant variables and a process that can discovering patterns from them to extract migration information with understandable structure should be made.

To analyze and make predictions for Chinese internal migrations, several facts might interfere the process. In despite of the difference of policies among provinces, the analysis of Chinese internal migration need to take the impact of policies into consideration. Super large cities like Shanghai and Beijing are implementing policies for controlling the total amount of migrators each year by enhancing the requirement for migrators to be registered as local residences, which is also called the Hukou (Residence Registration) policy. Some cities, however took actions in welcoming migrators from other provinces, especially migrators with high education background. Since Hukou policies play important role in migrations, an internal migration model in China need considering these policies. Understandingly, these policies are hard to be represented as numbers, which makes building models based on priori knowledge difficult in this area. To get rid of this problem, instead of using the data directly represent the policy impact, this paper utilizes economics and environmental data to indirectly represent it, which means, this paper is not using any direct policy data as priori knowledge for analyzing migrations but utilizing many other data that will be effected by the policy and, at the same time, highly related to migrations for modelling.

When considering the behavior of immigration, the “hidden residence” as mentioned by Robert (2002) is effective to the statistics analysis. Many Chinese citizens migrate for 1-year or even short time jobs but not for residence, this causes the real situation of migrators are hard to be represented by the data of local changes in registered residence. In this paper, with the contribution of National Bureau of Statistics of China, employs several investigated data sets from 2009 to 2014. These data are conducted by thousands of face-to-face surveys for migrators between 2009 to 2014, which can more clearly presented the true situations for individual migrator, especially for those without residence registrations.

In addition, the different cultures from various places inside China remain highly influent for migrations. A destination with a different culture, or with an accent that can not even be understandable for one migrator clearly cannot be attractable. The process of finding how different culture exact change the decision for one migrator could be unavailable since the description of one culture cannot be represented as numbers. In this case, the historical migration rates with distances between provinces can be used for representing the culture effect of one destination to migrators since with similar valuables in other attributes, the higher the migration rate is, the more similar culture they may share tends to.

This paper establishes models for analyzing migration attentions between provinces and individual intention for migration without much priori knowledge. Rather than using some attributes that has been considered as related to migration by priori knowledge for further analyzing, this paper use various attributes and only select few from them for further analysis by calculating and comparing the the maximal information coefficient scores before further classification algorithms and Artificial Neural Network (ANN) applied. Since one can select as many attributes as possible and the model will select potentially correlated features automatically, the process is built with little priori knowledge, which enable related problems in this area using the same process.

# 2. Design

Almost all meaningful analysis of internal migration in China begins with the analysis of the impact of various variables, such as economic development, policies, health-care and insurance onto the migration figures. However, in traditional methods of demographic analysis, to propose assumptions for the impact from several variables to internal migration requires much priori knowledge and will be time-consuming for researchers for analyzing each variables themselves.

To improve the process of internal migration analysis, a more automatic method is necessary. This method should be with the ability for selecting variables from received provided data, which are most correlated to target variable relation types, which is migration rate or the destination for one individual migrator each year in this paper, and be possible to reduce dimensions if necessary, in case of large amount of variables or data set for further reduction and, finally, be with a method for comparing various data mining algorithms for picking the best algorithm for a particular data set. This paper provides such an automatic process for migration analysis and use it into the Chinese internal migration research. The whole process contains the following particular algorithms.

Maximal information coefficient (MIC) is a measure of a wild range of associations both functional and not. A given internal migration dataset contains plenty of variables, which may contain important, undiscovered relationships between them. To compare correlations between pairs of data, a score for relationships between variables is necessary to be calculated and sorted for picking up important variables for further research. As a statistical method with the property of generality and equitability, MIC shows great performance in providing scores for correlations. As a general algorithm, MIC can capture a large range of relationships between variables based on the idea that if a concrete relationship exists between variables, then on the scatterplot, a grid of these two variables can be drawn that partitions the data to encapsulate that relationship (2).

The MIC score and strength relationships between variables can be shown as following:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Relationship | Random | Linear | Cubic | Exponential | Categorical | Parabolic | Sinusoidal |
| MIC | 0.18 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

A MIC score approximating to 1.00 denotes a very strong relationship between variable and a MIC score around 0.18 means the relationship is quite weak.

Principal component analysis (PCA) is a method for reducing dimensions for further analysis by converting a set of possible correlated data into a set of non-linearly correlated variables. The first few components calculated carries the larger variance (information). Then, the data will be projected into such components, which can maximize the mutual information between original and projected data. Taking a two-dimensional data for example, the direction is calculated to preserve the most information in the data set.



Img X

On image X, the observations marked by are projected onto a one-dimensional linear subspace with direction u, which can preserve largest possible information by one dimensional subspace.

More formally, for a data matrix X = [x1] … [xn]T, mean normalization should be done first for reducing the effects of different ranges of variables by computing the mean and replacing all data points xi with x̅ = xi – to result in a data covariance matrix **x̅**. Then, calculating the eigenvectors, which represent the variance of each dimension calculated by PCA, and the corresponding eigenvalues of the data covariance matrix . After obtaining all eigenvectors and eigenvalues, a decision of the total amount of dimensions M is required. To choose the principal subspace, simply choose the M largest eigenvalues and their associated eigenvectors, which are the basis of the principal subspace. Finally, the projected vector can be calculated by .

The Artificial Neural Networks (ANN) are composed of nodes or units connected by links and each link also has a numeric weight associated wit it, which determines the strength and sign of the connection (See img X) (3).



Img X

In this paper, only feed-forward networks have been deployed, which have connections only in one direction.

With perception training rule, a simplified one hidden layer neural network (perceptron) as shown in Img X can be trained in following steps:



Img X

1. Set weights randomly;
2. With one example come through, the output of the neural network can be calculated with the activation function and the weight between and the output layer by:
3. When an example is misclassified, change the weights with the following function using the weight between and the output layer , the learning rate and the training set with vectors with the corresponding targets :

After MIC and PCA, parameters can be selected and dimensions can be reduced automatically, which can reduce the requirement of priori knowledge, which used to be largely required by normal migration analysis.

In this paper, the resulted models are conducted in the following process.

The model for predicting migration rates between places:

1. Collecting data upon migration rates and other related variables as many as possible.
2. Calculating MIC scores between each variable and migration rate. Sorting the results and choose M most correlated variables for further analysis.
3. Using the M variables in various classification algorithms like Bagging, Linear Regression and Multilayer Perceptron. Choose the algorithm that with the largest RMS score.
4. Using the model from step 3 for further prediction.

The model for predicting migration destination of individual migrators.

1. Collecting data for individual migrators and their family situations and using the average GDP per capita in the destinations to represent them.
2. Employing PCA in the data set to reduce dimensions.
3. Running Artificial Neural Network (ANN) with the data after PCA
   1. Running ANN with the topology of 2 hidden layers with no more than 40 nodes in each layer to find the optimistic parameter set. In this process, 6-folder validation check is used for tuning the parameter configuration and the test data set for each topology is the whole original data set for ensuring the same test data is deployed through the process
   2. Using 4 different training functions in the ANN training process to find the best training function.
4. Predicting the destination of individual migrator by the nearest province or city with similar GDP per capita.

# 3. Data Description

Most data used in this paper is conducted by National Bureau of Statistics of China, the official agency under the State Council of China for collecting data related to the economy, population and society of the China at global and local levels (4).

There are basically three parts of data used in this paper.

The first set is the one that describes local economy, environment and population. This data set contains more than 20 dimensions, which can be collected from the database in National Bureau of Statistics. This data set illustrates the general economical, environmental circumstances of each provinces and direct-controlled municipalities, which are cities with status equal to provinces in China, which is similar to federal district. All the data in this data set is presented by the year of collected and the province or direct-controlled municipality it belongs to, which means the analysis related to this data set can only be regulated to the migration status between provinces and direct-controlled municipalities but not specified cities or counties. All data from this data set is publicized by National Bureau of Statistics of China

The second one is the unchanging geological data of provinces and direct-controlled municipalities. The most important dimension is the distances between provinces, which records the distance between the capital cities between each provinces and direct-controlled municipalities.

The last one is the interview data conducted by the National Bureau of Statistics collected from thousands of migration families. This huge interviewing project was taken each year from 2009 to 2014. The samples in this data set are selected by local officers, who also worked closely with migrators for asking questions as documented in the questionnaires and recording their answers. Total number of samples from each province or direct-controlled municipality are kept equal, which means the analysis upon this whole data can provides information towards the whole country without uncontrolled emphases on a certain geological area.

In despite of the detail on the questionnaires, more than 200 questions are listed. Some of them are optional questions that might not required to be recorded according to prior answers, which means the amount of missing data is not ignorable in this data set. For this reason, the particular missing data problem in this data set is different from normal missing data problem. Generally, missing data are caused by the misreporting in surveys but the missing data in this data set are data that should be missing according to the prior questions or the circumstances of the migrator’s families such like the whether the migrator has a partner or not. Since missing data under certain questions may carry some information, common methods like adding the mean of a certain column to the missing data can misrepresent the true information carried by missing data. Thus, in this paper, the missing data are added according to the questions of them, which will be further discussed later in this paper.

Since the questionnaires taken each year has slightly difference between each other, when selecting variables from the whole variable set for building models that not just suitable within one-year data, only common variables (data related to common questions) through out the 6 questionnaires within 5 years (2 different questionnaires are taken in 2010).

The form of data sets are shown in the following Img X. Due to the amount of parameters in Interview data, detailed columns of this data set is not shown below. Some of the columns of this data sets will be listed in the later chapters.



Img X

# 4 Prediction migration rates

As listed before, the whole process of this chapter can be divided into four steps. Among them, the step of selecting parameters and the process of building models using selected parameters with the data sets are most important.

In the process of building models, this paper employs several different methods and test the accuracy for each before finally choosing the most fitted model. This choosing process can make the idea of this chapter be suitable not only for data mine for Chinese internal migration but also for any related project.

Typically, in this analysis process, each time for training models, an iteration number of 10 has been chosen while in each iteration. Before training, the data will be divided into 10 parts equally. In each iteration, 9 out of the 10 parts are selected to be the training set with the one part left to be the testing set.

This method forces the same training method to be trained and tested by 10 times rather than once and calculate the final accuracy by average accuracy among test sets as shown in Img X.



Img X

All the results for prediction in this chapter are trained and tested with this process.

## 4.1 Migration Data

There are mainly three parts of data are used for training and testing in this chapter: General Information Data Set, Geography Info Data Set and Interview Data set.

In this chapter, the data sets need to be connected before further analysis. The final processed data set should contain variables about the two provinces that migrators come from and go to with the corresponding migration rates and other general information data of the provinces.

The relationship between data sets is shown in Img X.



Img X

Unlike other data that can directly obtain from three data set provided, the migration rates between provinces has not been directly recorded. Unfortunately, there is no official record about the real amount of migrations between provinces. If there are records of migration rates in some provinces, the large amount of unregistered migrators (so called “hidden migrators” ) would lead the records inaccurate.

In this circumstance, the migration rates used in this paper are calculated by the interview data sets. Since data in these sets are real data about migrators no matter whether they have been registered or not, using this data set to calculated the migration rate can provide a more closely migration rate to the real figure. However, there would be one problem exists in this process. Since records this data set is only sampling from the whole migrators, errors might exist, especially to these provinces with small migration rate, the rate calculated might close to 0.

## 4.2 Selecting features

For predicting the future immigration rate, features must be selected to build the model. During the process of evaluating the relevance between feature and target variable, MIC scores are used to select the valuables that contribute more influence on immigration rate than others.

The initial features can be selected from data in various areas, which considered might be correlated with the variable to be predicted. With the increasing amount of features selected, the finally picked parameters will become worse based on the method of picking highest MIC scores. This means, for any related topic, researchers should find as many features as they can that might influence the result.

In this paper, the listing parameters are selected for MIC sorting with migration rate:

|  |  |
| --- | --- |
| dis | Distances for immigration |
| orgGDP | The GDP per capita of the place immigrates from |
| Population | The population of the destiny |
| Birth\_Rate | Birth Rate of the destiny |
| Avg\_Life | Life expectancy |
| GDP\_PC | GDP per capita of the destiny |
| Unemploy\_Rate | Rate of unemployment |
| Flat\_Rates | Rates of flats |
| Price\_Rates | Rates of Price (last year: 100) |
| City\_Income | Average annual income for citizens live in the city area |
| Country\_income | Average annual income for citizens live in the rural area (Start publishing since 2013) |
| Green\_Rate | Rate of Green land in city area  (Start publishing since 2011) |
| Disasters | Amount of natural disasters happened in the area |
| Stu\_Rate\_H | Rate of citizens that obtains college certificates |
| Low\_Income | Rate of citizens that are classified as low income |
| Low\_Income\_Insurance | Monthly insurance for low income citizens |
| Flat\_Rates\_Change | Change in average rates of flats |

Before taking the all 5-year data for MIC scores, to be simplified, the data of 2011 is calculated and their MIC scores are shown below:

|  |  |  |
| --- | --- | --- |
| Y var | MIC (strength) | Linear regression (p) |
| Population | 0.67098 | 0.5396318 |
| Birth\_Rate | 0.67098 | -0.29075608 |
| Avg\_Life | 0.67098 | 0.4073089 |
| Flat\_Rates | 0.67098 | 0.53999525 |
| City\_Income | 0.67098 | 0.5662188 |
| Country\_income | 0.67098 | 0.5470502 |
| Green\_Rate | 0.67071 | 0.30561876 |
| GDP\_PC | 0.66989 | 0.48178324 |
| Disasters | 0.66989 | -0.075478196 |
| Stu\_Rate\_H | 0.66989 | 0.3403138 |
| Flat\_Rates\_Change | 0.63526 | -0.3043905 |
| Low\_Income\_Insurance | 0.62499 | 0.45970377 |
| Low\_Income | 0.61332 | -0.226193 |
| Price\_Rates | 0.37312 | -0.12579323 |
| Unemploy\_Rate | 0.21025 | -0.2366798 |
| orgGDP | 0.14091 | -0.014871608 |

This table with the MIC scores of the data in 2012 shows the features selected are mostly relevant to migration rate of 2011. As can be shown in the table, orgGDP, Unemploy\_Rate and Price\_Rate obtain lower MIC scores than others, which illustrates they uncorrelated with migration rate in 2012. Particularly, in this paper, a threshold is set to be 0.5 in MIC scores.

In this case, parameters except orgGDP, Unemploy\_Rate and Price\_Rate have been deleted for further analysis.

For further illustration about MIC and relationships between variables. A strong relationship between the amount of migration from each place to Beijing and the economic condition in their hometown is shown in Image1, in which the red line indicates the Highest GDP per capita value minus GDP per capita value and the blue line indicates the number of migrators and the total number of migrators has been used, which equals to migration ratio multiplied by total migrators from Beijing.

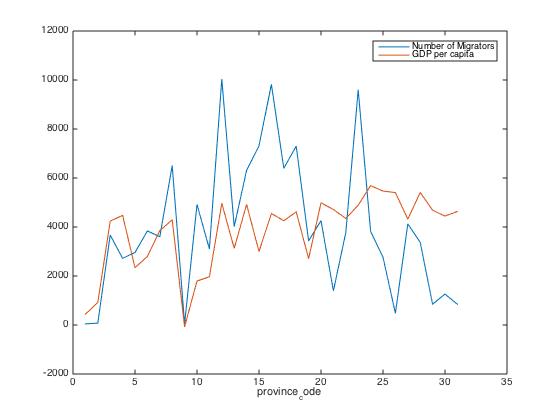


Image 1

As shown in Image 1, the provinces with fewer province codes follow the case that more rich (indicated by GDP per capita) one province is, more people will move out of the province, which, follows the migration model proposed by Harris and Todaro (6), which illustrates that the development economics is the important reason for rural-urban migration.

The list of Province names represented by Province Codes are shown in the following table.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Code | Province | Code | Province | Code | Province | Code | Province |
| 11 | Beijing | 31 | Shanghai | 42 | Hubei | 53 | Yunnan |
| 12 | Tianjiang | 32 | Jiangsu | 43 | Hunan | 54 | Xizang |
| 13 | Hebei | 33 | Zhejiang | 44 | Guangdong | 61 | Shanxi |
| 14 | Shanxi | 34 | Anhui | 45 | Guangxi | 62 | Gansu |
| 15 | Neimenggu | 35 | Fujian | 46 | Hainan | 63 | Qinhai |
| 21 | Liaoning | 36 | Jiangxi | 50 | Chongqing | 64 | Ningxia |
| 22 | Jilin | 37 | Shandong | 51 | Sichuan | 65 | Xinjiang |
| 23 | Heilongjiang | 41 | Henan | 52 | Guizhou | 53 | Yunnan |

Table1. Province codes

However, interestingly, provinces with large province code like 31 has very few people moves out while the GDP per capita in those provinces stays low. My assumption for this situation is that those provinces are located in far west part of China, which means their distance to major cities with high GDP per capita such as Beijing and Shanghai are quite far. Since this, people may feel unfavorable for migrating to richer areas typically, in east coast of China.

With comparing the MIC score, the scores for pairs of variables in the data set are not high enough to judge one single variable could strongly influence the others. However, among those pairs, the most relevant one is the pairs between the location migrators see the doctors with other features like monthly income and even the amount of children.

The scores of the relationship between Income and Number of Children are shown in table 1.

|  |  |
| --- | --- |
|  | Income-No. of Children |
| MIC | 0.00586 |
| MIC-P2 | 0.005460929 |
| Linear | 0.01997676 |

Table 1

Against the idea from most of the papers that the number of children of one family will be influenced by the monthly income of the family, the MIC score is just around 0.00586, which is one out of then when compared with the MIC scores with locations for migrators to see doctors.

In addition, relationships between the amount of people leaving the provinces with their location also be founded. Some provinces like *Anhui* (which is the province with the largest number of citizens migrate out) and *Sichuan* (province with second largest number of migrators leaving the province), also their GDP per capita are not quite low, however, since they locate very close to major cities and very rich areas in China like *Shanghai* (highest GDP per capita in China) and *Guangdong* (largest GDP in total), citizens in those provinces seem more intended to move to the province next to them and earn higher salary. The population of migrators from each province show as Image 2. (34: *Anhui*, 41: *Sichuan*)

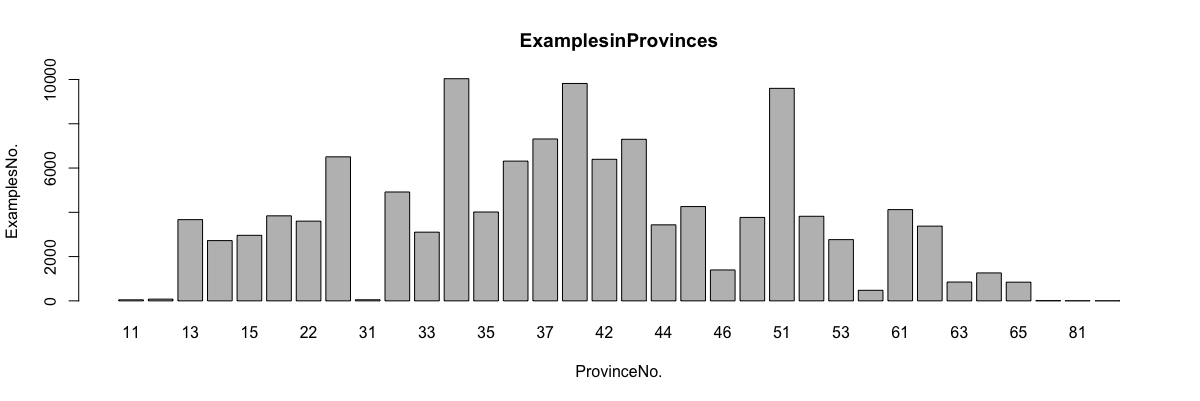
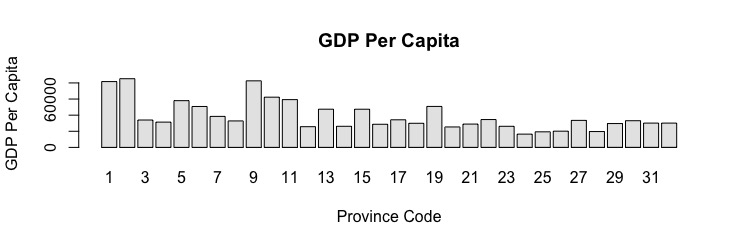
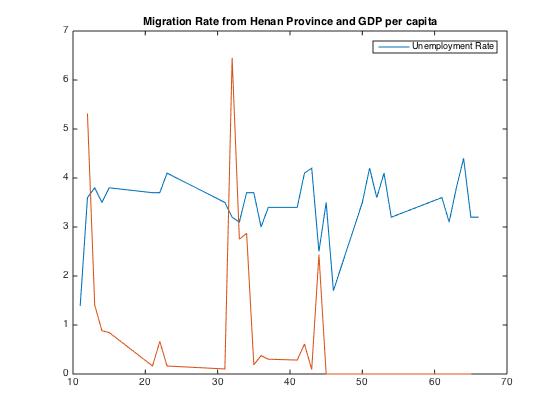


Image 2



For places like Hang Kong, Macau and Taiwan, since there is a border control between them and Mainland China, the migrators from those places are ignored in this project

In addition, taking “unemployment rate”, which has low MIC score for example. As can be judged from Img 3, the relationship between the migration rate from Henan and unemployment rate in destinations are not quite related.



IMG 3

One more detail that should attract attention is that, in this paper, general data of 2011 is used to make prediction of migration rate of 2012 rather than 2011. This make sense in China due to the culture of Spring Festival (also named as “Chinese New Year”), which is the most important festival all year round.

Spring Festival is celebrated at the first day of the traditional lunisolar Chinese calendar each year, which is a date between January and February. In Spring Festival, all Chinese people are supposed to travel back home for a holiday for 7 to around 20 days. Most Chinese people, especially migrators are leaving their job in cities and only move to seek job opportunities after the holiday. Normally, most companies who employ those migrators from rural places are supposed to pay the salary to employees from rural places according to their contract, which marks an end the contract and possibly an end to one-time migration. After the Spring Festival, these people leave their hometown again and head to the same or different big cities from last year for jobs. This tradition makes the Spring Festival become an end of migration and a start of a brand new migration.

Considering the impact of Spring Festival, when most migrators decide their destination for the upcoming year, they will depend on the data of years before rather than years in the future. Thus, only the analysis of the data in past few years make sense to this project.

One more parameter that has be deleted from the pre-selected parameters table is the Flat\_Rate\_Change due to its unusual performance. This deletion requires prior knowledge as following.

Generally, the change of flat rates should be positively related to the migration rate. This is because, generally speaking, the increasing amount of population in one city will cause the increasing of flat rate due to the increasing of need. The increase rate of flat price and its relationship with immigration rate is not a positive relation as most academies expected. The MIC shows there is a certain relation between them with a correlation score as 0.63626. However, if the data is analyzed with linear model the score is -0.3043905, which denotes that, generally, there is a certain negative relation between. This might due to the influence from politics, since the government take actions for reducing the high flat rate in big cities for improving the living environment for locals.

Another reason for deleting this parameter is that, the upcoming years after 2012, the relationship between changes of flat rates and migration rate turns to be positive. This may also due to the unpredicted policy changes within the upcoming years. Thus, taking this parameter for further analysis will mislead the models and cause over fitting problem.

With the same MIC calculator, a data set with all statistics data from 2009 to 2013 is used to predict the immigration rate in 2014 between provinces. There relevance to immigration rate is as follow:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | variables | MIC Score | Rank | variabes | MIC Score |
| 1 | Birth\_Rate\_13 | 0.67098 | 33 | Low\_Income\_Insurance\_13 | 0.62499 |
| 2 | Birth\_Rate\_10 | 0.67098 | 34 | Low\_Income\_Insurance\_12 | 0.62499 |
| 3 | Avg\_Life\_13 | 0.67098 | 35 | Low\_Income\_Insurance\_11 | 0.62499 |
| 4 | Avg\_Life\_12 | 0.67098 | 36 | Low\_Income\_Insurance\_10 | 0.62499 |
| 5 | Avg\_Life\_11 | 0.67098 | 37 | Low\_Income\_Insurance\_09 | 0.62499 |
| 6 | Avg\_Life\_10 | 0.67098 | 38 | Birth\_Rate\_13 | 0.62499 |
| 7 | Avg\_Life\_09 | 0.67098 | 39 | Birth\_Rate\_12 | 0.61807 |
| 8 | GDP\_PC\_09 | 0.67098 | 40 | Birth\_Rate\_11 | 0.61807 |
| 9 | City\_Income\_13 | 0.67098 | 41 | Green\_Rate\_09 | 0.61807 |
| 10 | City\_Income\_12 | 0.67098 | 42 | Green\_Rate\_11 | 0.61332 |
| 11 | City\_Income\_11 | 0.67098 | 43 | Low\_Income\_13 | 0.61332 |
| 12 | City\_Income\_10 | 0.67098 | 44 | Low\_Income\_12 | 0.61332 |
| 13 | City\_Income\_09 | 0.67098 | 45 | Low\_Income\_11 | 0.61332 |
| 14 | Green\_Rate\_13 | 0.67011 | 46 | Low\_Income\_10 | 0.61332 |
| 15 | GDP\_PC\_13 | 0.66989 | 47 | Low\_Income\_09 | 0.61332 |
| 16 | GDP\_PC\_12 | 0.66989 | 48 | Birth\_Rate\_09 | 0.58389 |
| 17 | GDP\_PC\_11 | 0.66989 | 49 | Green\_Rate\_10 | 0.56935 |
| 18 | GDP\_PC\_10 | 0.66989 | 50 | Green\_Rate\_12 | 0.56422 |
| 19 | Stu\_Rate\_H\_13 | 0.66989 | 51 | Unemploy\_Rate\_12 | 0.52076 |
| 20 | Stu\_Rate\_H\_12 | 0.66989 | 52 | Unemploy\_Rate\_11 | 0.52015 |
| 21 | Stu\_Rate\_H\_11 | 0.66989 | 53 | Disasters\_09 | 0.45272 |
| 22 | Stu\_Rate\_H\_10 | 0.66989 | 54 | Unemploy\_Rate\_13 | 0.43731 |
| 23 | Stu\_Rate\_H\_09 | 0.66989 | 55 | Disasters\_10 | 0.42073 |
| 24 | Population\_13 | 0.66936 | 56 | Price\_Rates\_09 | 0.41893 |
| 25 | Population\_12 | 0.66936 | 57 | Disasters\_12 | 0.3995 |
| 26 | Population\_11 | 0.66936 | 58 | Price\_Rates\_13 | 0.39465 |
| 27 | Population\_10 | 0.66936 | 59 | Disasters\_11 | 0.38736 |
| 28 | Population\_09 | 0.66936 | 60 | Price\_Rates\_09 | 0.37312 |
| 29 | Country\_income\_13 | 0.66878 | 61 | Price\_Rates\_12 | 0.32843 |
| 30 | Country\_income\_12 | 0.66878 | 62 | Unemploy\_Rate\_09 | 0.22774 |
| 31 | Country\_income\_11 | 0.66878 | 63 | Unemploy\_Rate\_10 | 0.21025 |
| 32 | Country\_income\_10 | 0.66878 | 64 | Price\_Rates\_11 | 0.20614 |
| 33 | Country\_income\_09 | 0.66878 | 65 | orgGDP | 0.14091 |

Interestingly, two phenomena can be found from this table. Parameters with similar meanings in different years mostly obtain similar ranks and for these parameters, in more recently they are collected, the more correlated they are to the migration rate.

## 4.3 Building the model:

As mentioned before, 6 methods are used in accuracy comparison in this section, they are: Bagging, Multilayer Perceptron, Additive Regression, REP Tree, Decision Stump and Linear Regression.

### 4.3.1 Bagging

Bagging, short for Bootstrap Aggregating, is the method to generate several predictors from selecting data into different subsets randomly for several times and combine them into an aggregated predictor. To combine predictors, the aggregation calculates the mean over versions of predictors when making prediction.



Img X

Based on the idea that forming multiple versions of predictors by making bootstrap replicates of the whole learning set and using these as new learning sets. Bagging can give substantial gains in accuracy as proved by Leo Breiman in 1996 (5).

Based on the subsets generated in bagging aggregation, decision trees are built to classify the data and make predictions to further input.

Decision Tree (DTree) is a support tool using a dendritic graph or model of decisions and their possible consequences (7). It is a tree with leaf nodes, which has a class label, determined by majority vote of training examples reaching the leaf and internal nodes, which carry questions of features.

In this paper, C4.5 is used for building the trees after bagging aggregations. C4.5 is a method for choosing features in each internal nodes, which is similar to the common used ID3 method. The difference happens when deciding the chosen feathers in each inter node after the calculation of information gain. Instead of using information gain as the quartier for feather choosing, the C4.5 uses a information ratio as quainter. The C4.5 defines split information with “D” for total amount of examples in this node and for the amount of examples with feature ‘A’ valued as ‘I’ and n stands for the total kinds of values in this feature :

After obtaining the split information, the information gain ratio can be calculated as:

The value of is calculated with the same way in ID3, which is:

The process of this algorithm can be shown in Img X

With the algorithms as mentioned before, the statistics data of 2010 is used below in building models for migration rate in 2011 to show the accuracy of each learning methods towards the data.

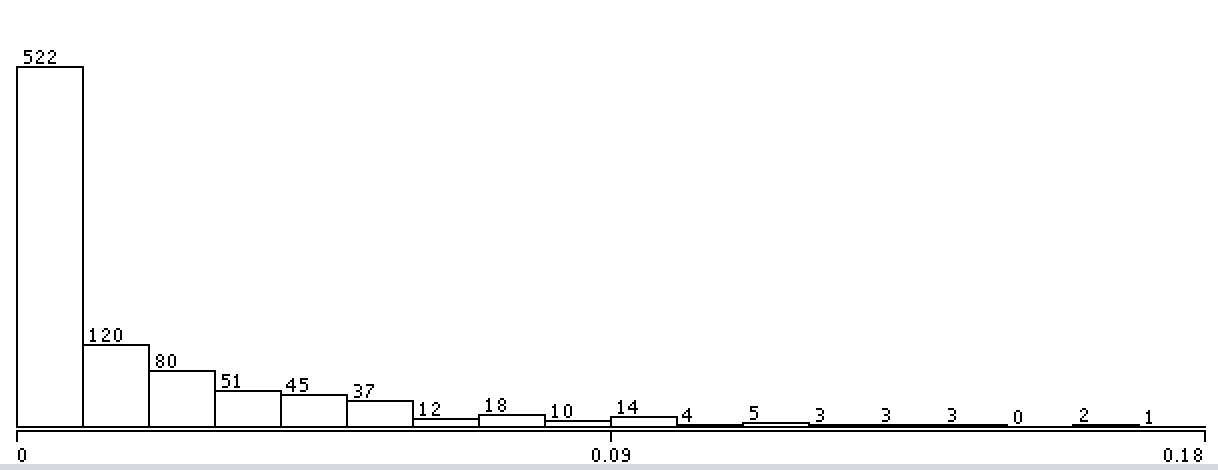
The accuracy of each training method with the use of RMS(Root Mean Square) to comparing between different classifiers for predicting Immigration Rate is shown blow:

|  |  |  |
| --- | --- | --- |
| Method | RMS | Mean absolute error |
| Bagging | 0.0404 | 0.017 |

Using this method for making prediction with statistics data of 2010 for migration to 2012 and compare them from the migration rates collected and calculated from the interview data set. The results will be shown in the following format:

|  |  |  |
| --- | --- | --- |
| Actual | Prediction | Error |
| 0 | 0.001 | 0.001 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0.015 | 0.076 | 0.062 |
| 0.006 | 0.007 | 0.001 |
| 0.001 | 0.016 | 0.015 |
| 0 | 0 | 0 |
| 0.049 | 0.097 | 0.048 |
| 0.032 | 0.02 | -0.011 |
| 0.001 | 0.01 | 0.009 |
| 0.167 | 0.166 | -0.001 |
| 0 | 0.03 | 0.03 |
| 0.004 | 0.009 | 0.005 |

Error Bar:



With errors as follow:

Mean absolute error 0.0225

Root mean squared error 0.0604

As shown above, the Bagging aggregation with C4.5 makes prediction for 2 years with high precision with only around 0.05 in RMS. This method can provide predictions for future migration rates between provinces for policy makers.

To find the best method for building models for internal migration, several more algorithms are also employed.

### 4.3.2 Linear Regression

Linear Regression is also selected as potential algorithm for building models for internal migration in China since most variables listed in the chapter for MIC calculation also obtains positive scores in linear regression. Especially, variables like “city\_income” and “population” seem to have a relative strong positive correlation with migration rates.

As a model that assumes all variables follow linear relation towards the target variable , linear regression can be expressed in the following equation with the denotes the small and acceptable noise.

More generally, for a set of samples with values and target values. The model takes the form

To be simplified, set and . Using ordinary least squares, which is a common estimator that reduce the total amount of squared residuals, the equation for the estimated value of the parameter set , which is short for can be expressed as

Employing linear regression to the 2010 statistics data and 2011 migration rates, the RMS and mean absolute error are shown below

|  |  |  |
| --- | --- | --- |
| Method | RMS | Mean absolute error |
| Linear Regression | 0.045 | 0.0215 |

Using this method for making prediction with statistics data of 2010 for migration to 2012 and compare them from the migration rates collected and calculated from the interview data set. The erros are:

Mean absolute error 0.0261

Root mean squared error 0.067

This is slightly worse than Bagging.

### 4.3.3 Random Forest

Similar to “bagging”, which has been mentioned in chapter 4.3.1, Random Forest also uses the idea for splitting data sets for training. Moreover, Random Forest does also selects feathers while building the trees for each subsets.

Random forest was proposed by Breiman in 2001 as an extension for “bagging” (8). As described by Breiman, random forests are a group of decision trees such that each individual tree replies on the values of a random vector, which is sampled independently and with the same distribution for all trees in the forest (9).

Unlike the algorithm described under chapter 4.3.1, random forests change the way that decision trees are constructed. In standard decision trees, the node splitting methods use information of the node for finding the best splitting decision among all variables in the data set. However, in building decision forests, only the best feathers among the subset of predictors, which are randomly chosen over the nodes can be used for splitting the nodes. This strategy turns out to obtain a very well performance among many other classifiers as mentioned by Breiman in 2011 (8).

The algorithms follow the steps below:

1. Using bagging for selecting subsets from the original data;
2. For each subset that generated by the bagging, generate a decision tree ( using C4.5 in this paper). During the processing of generating, at each internal node, rather than choosing the best features among all predictors as described in chapter 3.3.1, selecting a random sample of the features and only use information gain ratio to select the best feature among them.
3. In the process of prediction after building the random forest, predict by aggregating the predictions made by the generated trees.

Specifically, in the following building process, the size of each bag and total number of iterations are all set to be 100.

The whole process is shown in Img X



Img X

As shown in the graph, if the subsets of features are set to be the same as the whole feature set, random forests will be the same as bagging aggregating, which means that bagging can be taken as a special case of the algorithm of random forests.

After applying the algorithm of random forests on statistics data of 2010 and migration rates in 2011, the results are:

|  |  |  |
| --- | --- | --- |
| Method | RMS | Mean absolute error |
| Random Forests | 0.0397 | 0.0169 |

As shown in the table, in this data set, random forests have a better performance over bagging.

Following the steps of prior algorithms, the prediction accuracy on real data using the model just built based on statistics data of 2010 and migration rates in 2011 to predict the migration rates in 2012. The result is:

Mean absolute error 0.0201

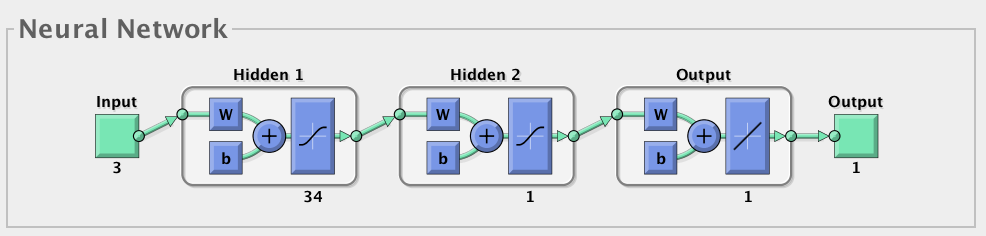
Root mean squared error 0.0572

The result shows the random forests just built perform slightly better than bagging.

* + 1. Feedforward Neural Network

The feedforward neural network is a kind of neural network, which obtains no cycle among nodes.

In this chapter, a multi-layer perceptron is used for predicting the migration rates between provinces. A simple 2 layer perceptron can be shown as Img X.



Img X

As shown in the map, the number of nodes in each layer of the neural network need to be set specifically. Selecting the number of nodes in each hidden layer can directly interfere the accuracy of the whole model. However, up to now, there is no universal method for deciding the number of nodes in each layer before running the algorithm.

In this paper, a method of running the neural network with the combination of 1 to 40 nodes in hidden layers and evaluating the corresponding accuracy has ben employed in this paper to make judgment of the best number of nodes in each layer.

A library in Matlab is used for building neural networks in this chapter. With the use of this library, several different multilayer neural network algorithms have been employed. As to find the best algorithm fitted for Chinese internal migration, 4 neural network methods has been run one by one for each combination of 1 to 40 nodes in hidden layers.

To equally verified the accuracy for each iteration and each neural network method, since Matlab will automatically and randomly divide the data set into training set and test set before each iteration, which will definitely cause the difference in accuracy estimation even for the same method and number of nodes in each layer. This paper uses the whole data set as testing set, which controls the test set to be the same for all four different algorithms and for all iterations.

The four neural network algorithms are:

1. Gradient descent backpropagation (traingd) – Parameter: learning rate (lr).
2. Gradient descent with adaptive learning rate backpropagation (traingda) – Parameters: learning rate (lr), ratio increase/decrease learning rate (lr\_inc, lr\_dec).
3. Gradient descent with momentum backpropagation (traingdm) – Parameters: learning rate (lr), momentum constant (mc).
4. Resilient backpropagation (trainrp) – Parameters: Increment/Decrement to weight change (delt\_inc/delt\_dec).

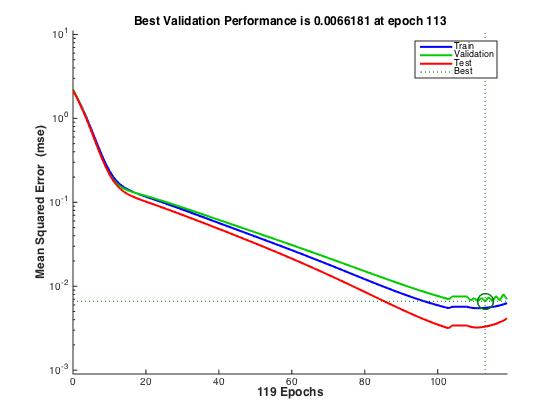
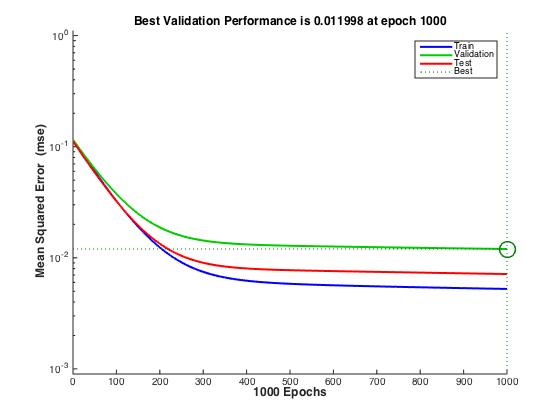
To run four method in Matlab, simply set the network by following steps:

1. Set number of nodes in each layer: “net = feedforwardnet([i,j]);”
2. Set maximum training episodes by ” net.trainParam.epochs = 1000;”
3. Configure the neural network by “net = configure(net, Input, Output);”
4. Train the network by “net = trainrp(net, Input, Output);”
5. Test the neural network by “y = net(Input); p = perform(net,Output,y);” with the whole data set

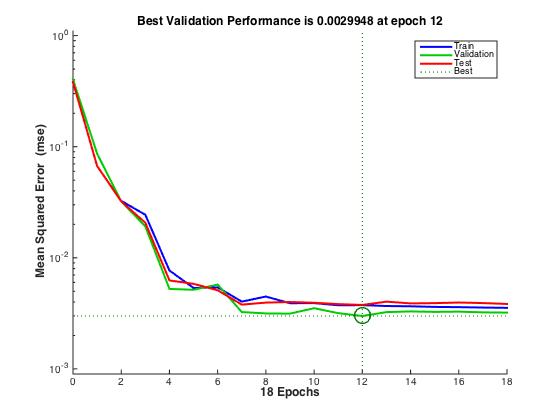
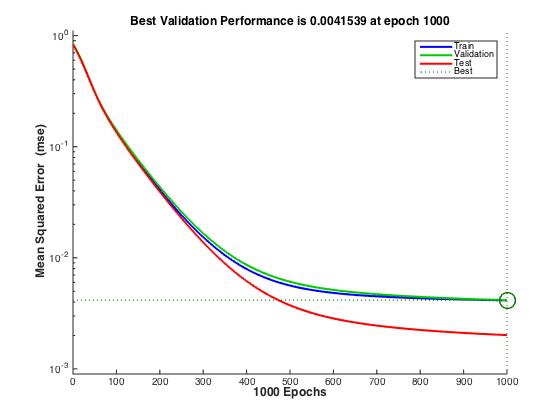
After setting maximum epochs as 1000 and training and testing all 1600 \* 4 neural networks in Matlab, the result of using statistics data in 2010 and migration rate in 2011 is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| Method | RMS | Nodes in 1st layer | Nodes in 2nd layer |
| Gradient descent backpropagation | 0.07 | 1 | 2 |
| Gradient descent with adaptive learning rate backpropagation | 0.0583 | 7 | 17 |
| Gradient descent with momentum backpropagation | 0.007 | 7 | 1 |
| Resilient backpropagation | 0.045 | 10 | 2 |

With the regression as follow:



traingd traingda



traingdm traingrp

One point needed to be mentioned is that Gradient descent backpropagation and Gradient descent with momentum backpropagation has not reached the regression at 1000 epochs.

After using the 2010 data for 2012 migration rates, the result is:

RMS: 0.0624

## Future Prediction

All the data mentioned in the prior chapter is the result of 1 or 2 year prediction using only statistics data in one year. These results can reflect the accuracy and efficiency of each method used before. In this fast developing society of China, using more history data for future prediction or making predictions for future 5 more years become more meaningful.

As mentioned before in the chapter of calculating MIC scores, the potential variables for model building and predictions making has been posted. Based on the data within that table, a model that based on data in more than one years can be used in building models.

Since the data set mentioned before contains only data between 2009 to 2013. Using migration rate of 2014, which can be calculated by the interview data set of 2014, the comparison can be made by building models with same algorithms based on data from 2009 to 2013 and with the data of 2013 alone, to check the influence of the amount of history data.

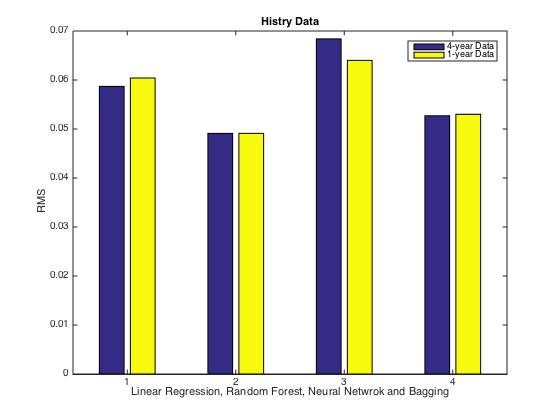
After building the models with the methods mentioned above, the accuracy of each method can be calculated as follow

|  |  |  |
| --- | --- | --- |
| Method | RMS | Mean absolute error |
| Linear Regression | 0.0587 | 0.0223 |
| Random Forest | 0.0491 | 0.0185 |
| Neural Network | 0.0684 | 0.0385 |
| Bagging | 0.0527 | 0.0197 |

And with the data of 2013, the result is as following

|  |  |  |
| --- | --- | --- |
| Method | RMS | Mean absolute error |
| Linear Regression | 0.0604 | 0.0277 |
| Random Forest | 0.0491 | 0.0185 |
| Neural Network | 0.064 | 0.0383 |
| Bagging | 0.053 | 0.0198 |

From these figures, predictions based on statistics data for more than one years can improve the performance of linear regression. However, the improvement on random forest and bagging is so slightly that can be ignored. Neither, for feedforward neural network, the data for more than one year cast negative effect on the precision of prediction as the shown in Img X.



Img X

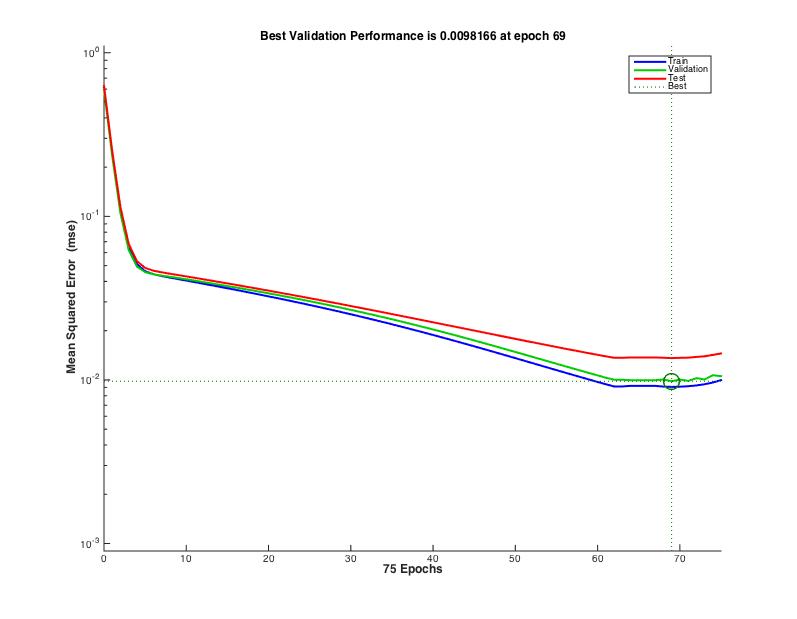
This is an unusual phenomenon since most theories proposes that in internal migration, more ages of statistics data employed in migration or relevant problem should improve the performance. More generously, most theories upon analysis on migration encourage researchers for applying more data into models to help the algorithms in learning potential information and relationships between variables. However, the data comparison shows that for some methods in this project, always employing a larger amount of data seems not a robust method in improving the performance.

There are two reasons for this problem.

1. Values of each variable scattered in each years for the same place show no distinct changes among years. As shown in the MIC scores, the relevance of most variables in various years in same area shows tiny differences between each other, which means, using the most recent value of the variable can about to represent the variable in each years. One more phenomenon to support this is that, for most variables, the most recent value (2013 in this case) obtains the highest MIC scores compared with the same variable in all other years, this illustrates that the most recent value be slightly more relevant to migration rate and values in other years may contain more irrelevant values that may interfere the result;
2. Over-fitting problem will occur when passing too much data.

For the second reason mentioned above, random forest and bagging contains ideology of preventing over-fitting by cutting of training set or cutting of features. Thus, for these methods, passing more features will be hard to interfere the accuracy to the already regressed model.

For the feedforward neural networks, adding data will cause an increase in validating accuracy, however, when testing with the test set although they might have already been over-fitting. Thus, when testing the model, the over-fitted model will behave worse than expected. For example, the image below shows an over-fitting neural network’s training process, the best parameters for validating are not the same as for testing set.



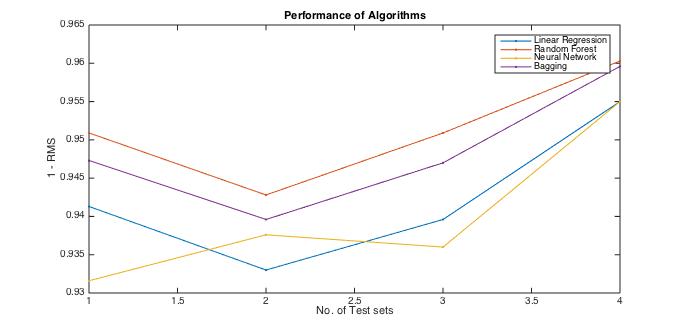
ImgX

As shown in the ImgX, the weights set at 69 epochs show the best among performance using validating data, however, while tested by testing data, it is slightly worse than the performance of the weight set around 63.

## 4.5 Evaluation

This chapter enhances four different kind of algorithms and with four different propagation methods for feedforward artificial neural networks in the process of migration rate estimation. The prior process has built models based on several sets of statistics data in China with these algorithms and compared the performance among them.

Based on the results shown in this chapter, random forest stays on top on the accuracy rank of all four kind of algorithms, with slightly higher accuracy than Bagging. The average RMS is shown in the following graph.



Img X

As shown in the graph, random forest obtains best accuracy can always perform better than other algorithms with bagging following it.

One common feature of random forest and bagging is that, both of these algorithms employ an idea of splitting for increasing accuracy and forbidding over-fitting. The features in the statistics data are not independent, in fact, according to the table below, which show an example of correlations between variables and “GDP\_PC” in the year of 2011, most features are highly correlated between each other.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | MIC | Variable | MIC |
| Disasters | 0.9769869 | Stu\_Rate\_H | 0.42304355 |
| Price\_Rates | 0.9338638 | Population | 0.3854314 |
| Unemploy\_Rate | 0.9103233 | Flat\_Rates | 0.3843422 |
| Green\_Rate | 0.88238555 | Avg\_Life | 0.37655085 |
| Low\_Income | 0.83737725 | Low\_Income\_Insurance | 0.22350341 |
| Flat\_Rates\_Change | 0.78134805 | City\_Income | 0.2158271 |
| Birth\_Rate | 0.60136974 | Country\_income | 0.15467566 |
| amount | 0.5621874 | fromCode | 0.001148369 |

With correlated variables, the model could be easily over-fitted by training on much data that not carries much difference between them. This will cause the model fitting the noise that exist in the data rather than learning the hidden information.

Thus, when researching social topics that includes feathers highly rely on others, it is important for not to use too much data with similar meaning for building models.

The general performance of these methods show positive result towards the prediction of migration rates. As shown in Img X above, RMS for most algorithms are around 0.05.

Pacifically, for prediction the migration rates in China, random forest is the best in performance.

The process of migration rate prediction thus can be listed below

1. Preparing, combining data sets with migration rates, economical and environmental variables in different social area. For best performance, migration rates data should be one year earlier than other statistics data;
2. Calculating and ranking MIC scores for each variables towards the target variable and selecting the variables based on their scores;
3. Training and testing several training methods includes neural networks, linear regression, random forest and bagging for choosing the best method over prediction with one-year internal migration data;
4. Using the algorithm with the best result over internal migration and use the model for data in other years.

# 5. Individual prediction:

Chapter 4 illustrates a process that achieve a relatively high accuracy over Chinese internal migration rates. The prediction over internal migration rates contribute to the overall general inspect over situations about the migration data. However, figuring out what are the reason for leading the individual migrators for choosing their own destinations is vital for understanding what would be the key reason for their migrating and for analyzing their willing of travelling based on individual data.

In this chapter, methods that for making predictions about the potential destination of each family of migrator will be illustrated and tested. Since in the interview data, which is the only concrete source of data concerning the real situation of Chinese internal migration, all interviewers were recording the migrators along with the information of their families into one records, the data of an individual migrator in this chapter also carries the information about family members of that individual.

For making prediction towards the individuals, a further notice about the data set of the interview data should be further illustrated since unlike other data sets used in the prior chapter, these data sets contain problems like considerable large amount of missing data, nonnumeric values in variable and larger size of data.

Besides of the theses problems in the data sets, the represent of destination in the data base will also cause big influences towards the accuracy due to all destinations are shown with province codes. This paper uses the GDP per capita of the destinations for representing the province codes with the reason that will be illustrated in the following chapter. This representation actually changes the aim of algorithms in this model from predicting the exact province to predicting the economic environment one actually would like to move to. After the prediction of GDP per capita, further analysis based on the geographic location for prediction of the real potential destinations can be made. Thus, the process can be described as follow

1. Data preparation for generating data sets that can be used for further analysis and replace the province code (both the origin and destination) with local GDP per capita;
2. Decreasing the dimensions of the data;
3. Training and testing the model, employing algorithms for predicting the GDP per capita that one migrator would prefer to migrate to;
4. If necessary, list provinces that locates near one’s origin with the similar GDP per capita as the predicted GDP per capita of the destination.

Thus, the result of this model is either a predicted GDP per capita that one migrator likely to travel to or a list of potential provinces but not just one province for the migrator.

## 5.1 Data Preparation:

The over 200 parameters are listed in the interview data set and the differences in questionnaires among years create difficulties for even just selecting variables that exists in every questionnaire. For example, on the questionnaire of 2013, there is a question for whoever or not the migrator holds a residence pass (question number: q202), which is a pass that provide migrators, who have not registered as a permanent residence in some cities with same welfare policy as local permanent residence. However, on questionnaires before 2013, no such question is listed on them since there was no such a policy for residence pass until 2013.

After deselecting questions that related that contains parameters that not exist in every interview data sets, those data with values of nonnumeric values should also be replicated or deleted.

The most important nonnumeric parameter in the questionnaires is about the place that the migrator is now living in. Although all questionnaires contain such a question but not all data sets record the place by its province codes. Some questionnaires like the one taken in 2012, record the Chinese characters rather than province codes.

For these records, a further action is needed to replace the characters with province code as used in other data sets. However, since the destination province is the target variable in the model of individual migrator prediction, simply just use province codes in the process of predicting will cause the increase of errors since codes fall in carrying much internal information or integrity for provinces considering that few provinces with vary different economic environment and geographic localization may have very close codes.

For this reason, GDP Per Capita is used in presenting different provinces. In the process of prediction, the Neural Network will output predicted GDP Per Capita of the destination that one family likely intended to travel to. If further prediction requires to know exactly the place rather than the GDP Per Capita of the destination, a province that near the original province which obtains close GDP Per Capita of the prediction will be chosen to be the place one probably travels to.

After filtering the data set following steps above and deleting parameters that contains too much Null values, 43 feathers are used to represent the data. Most of them are from the interviews taken by Chinese government. Unfortunately, among the 43 feathers, 28 of them contains Null value. So, before further analysis, missing data must be imputation.

Multiple methods have been created by researchers for the imputation of missing data like multiple imputation and partial imputation. However, most of them are designed for data that missing at random or even missing completely at random, which is not the case in this project.

As mentioned in prior chapters, most of them are caused by the designing of the questionnaire, which requiring some of people only take parts of questions. For example, in question marked by q204 of questionnaire taken in 2012 that asking for the type of job the migrator currently taking. Some migrators do not have a job while at the time of the interview and there is no choice for “not having a job”. Thus, such an answer will be recorded empty and finally be taken as a missing data in this project.

To back in this point of view, several parameters are listed below for valuing how much data are missing due the reason and how much are really missing randomly. The table below records the amount of missing records in parameters.

|  |  |  |
| --- | --- | --- |
| Parameter code | Number of missing data | Asking about |
| Q203 | 9531 | The industry of one migrator |
| Q204 | 9531 | The kind of job |
| Q206 | 9531 | The role |
| Q207 | 9531 | The average working days |
| Q208 | 9531 | The average working hours |

Along with records in q202, which asking whether the migrators have jobs:

|  |  |  |
| --- | --- | --- |
| Choices | Amount of selected | Total amount |
| Being employed | 55968 | 55968 |
| Lost the job | 866 | 9531 |
| Have not found a Job | 2690 |
| Not willing to take a job | 5777 |
| Retired | 198 |

As shown in the tables, a clearly conclusion can be made that most the missing data in these parameters have no jobs during that time. Based on this conclusion, we can frankly extend the conclusion to all missing data in the interview data sets, since the data collected are in good condition, which has been shown by above tables since there is no even one mismatch between the amount of unemployed migrators and the amount of missing data in these parameters.

This paper proposes various default values for feathers valued as Null according to the questions themselves, which means different questions will have different default values.

Methods for Missing Data imputation:

For data missing in question q101a2 to q101a6, which requests information about whether immigrates have certain family members travelling with them, a method simply refill the blank with 0 has been used in this project.

For records under question q203, which asking about the area that immigrates working in, a number 15 denoting that the immigrate worked in another unmentioned area.

For records under question q204, which asking about the kind of job that immigrates working in, a record 80 denoting that the immigrate worked in another unmentioned area is used in filling the blanks.

For records under question q205, which asking about the kind of cooperation that immigrates working in, a record 12 denoting that the immigrate did not work in any kind of cooperation as mentioned under the question is used in filling the blanks.

For records under question q206, which asking about the role inside the cooperation that immigrates working in, a record 12 denoting that the immigrate did not work as any kind of role as mentioned under the question is used in filling the blanks.

For records under question q207, which asking about the income last month for the immigrate, a record 0 denoting that the immigrate did not have any income is used in filling the blanks.

For records under question q208, which asking about the average working days for the immigrate last month, a record 0 denoting that the immigrate did not work last month is used in filling the blanks.

For records under question q209, which asking about the average working hours per day for the immigrate last month, a record 0 denoting that the immigrate did not work last month is used in filling the blanks.

For records under question q301, which asking about the time for the wedding of the immigrate, a record 201300 denoting that the immigrate has not married until 2012 is used in filling the blanks.

For records under question q304, which asking about the whether the family obtained the prof for one child family, a record 0 denoting that the immigrate has no child until 2012 is used in filling the blanks.

For records under question q306, q308, q309, q310, which ask about the whether the sexual life about migrators, a record 0 for q306 to q309 and a record 9 for q310 denoting that the immigrate has no sexual life during the last month for taking the questionnaire is used in filling the blanks.

For records under question q311, which asks the times of abortion made in recent 1 year, a record 0 is added the the blanks.

For records under question q312, q313, q314, q315, q316, which ask if the migrator or his or her partner has signed or obtains some documents for their children in the name of birth control. A record 0 is added to fill the blanks, which indicates that the migrator is neither signed nor required to sign nor cannot remember because the migrator do not need to sign it.

After appending these data, the data sets can be fulfilled properly and will be ready for further analysis.

## 5.1 Methods:

Unlike the prior process of building the model described, the MIC scores do not show much meaning in this section. After calculating the MIC scores, a very interesting point has been found that almost all parameters show very low correlation towards the destination of migrators.

The reason of this phenomenon should lay on the fact that single parameter of one record cannot determine the destination from a migrator. More generally, there is no common reason for every one in China that can determine the destination of one migrator.

According to the algorithm of MIC, only parameters, whose values as below that can be drawn in small grids on scatterplot can achieve high scores.



Unfortunately, since there is no dominant variable in this data set, sorting MIC scores seem to be meaningless and even misleading since almost all parameters show very low correlation to target variable on their own.

For this reason, the following process should proceed in analyzing all parameters in the data set, which makes the data sets relatively huge and high dimensional.

Since there is no direct relation between variables, algorithms like random forest, bagging and linear regression can hardly show high accuracy. For this reason, the following chapter focus on using artificial neural network for making predictions.

As mentioned before, the most important parameters for feed forward artificial neural network lay on the number of nodes in each layer and the algorithm of propagating in each term.

In the process of training the model for predicting potential destinations for individuals, to find the number of nodes in each layers for best result, a training process, which running neural network for 40\*40 times are used to iterate every possible combination of layers.

However, unlike the data sets used in the prediction of migration rate, the data set used in this chapter contains larger amount of samples and parameters. With 43 dimensions of data, running Neural Network through out all the data with dimensions from 1 to 40 for less than 3 layers and 4 propagating algorithms would be time-expensive.

For this reason, a method that can reduce the dimensions and keep the variance within parameters need to be used before running ANN in these datasets.

## 5.3 Dimension Reduction

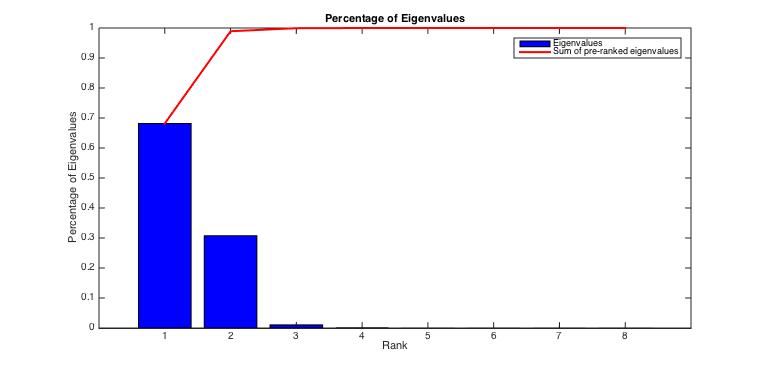
This paper uses principle component analysis for the reducing in dimensions in the refined interviewing data sets.

As mentioned in prior chapters, PCA calculates the eigenvalues and eigenvectors of and ranking eigenvalues for finding the principle component with the largest possible variance.

After mean normalization and the calculation of eigenvalues and eigenvectors over the interviewing data set of 2012 after imputing missing data with the method above, 46 eigenvectors and their corresponded are obtained. Among them, the top ten eigenvalues are listed below:

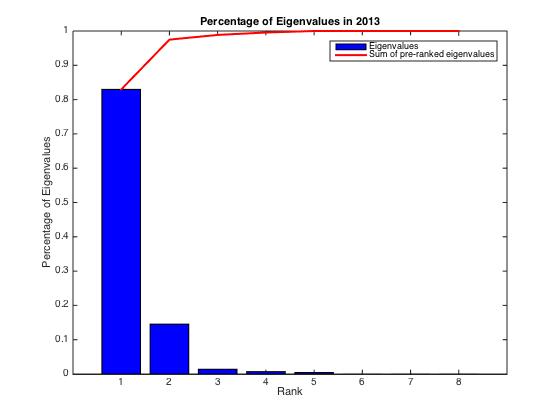
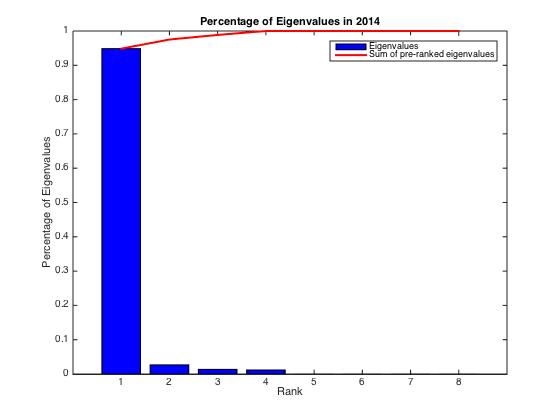
|  |  |
| --- | --- |
| Rank | Eigenvalues |
| 1 | 10784886531.5479 |
| 2 | 4863993192.10038 |
| 3 | 161469356.162105 |
| 4 | 9666079.31238827 |
| 5 | 559689.639788497 |
| 6 | 252.512592491710 |
| 7 | 110.774335783614 |
| 8 | 40.6530903552938 |
| 9 | 29.8874518810412 |
| 10 | 10.6164523524327 |

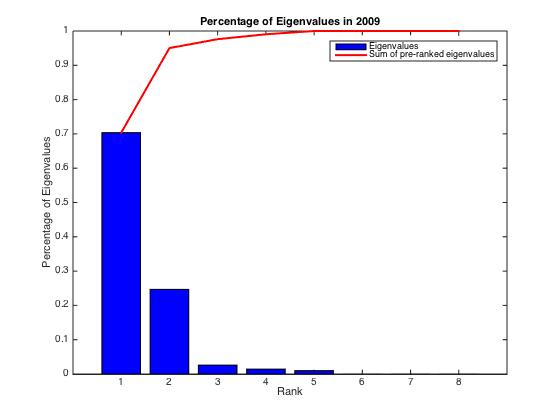
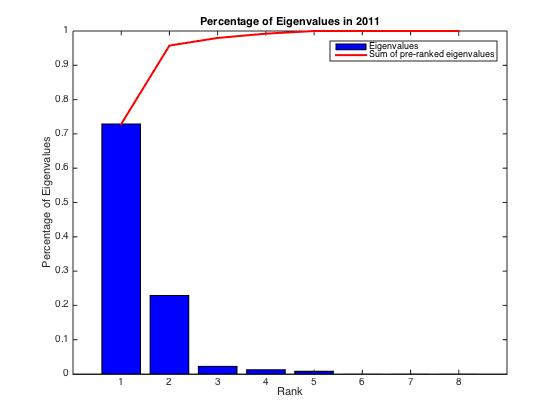
The percentage of each eigenvalues towards the sum of eigenvalues are shown below

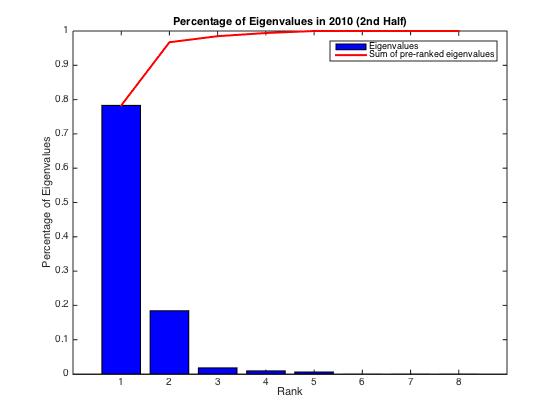
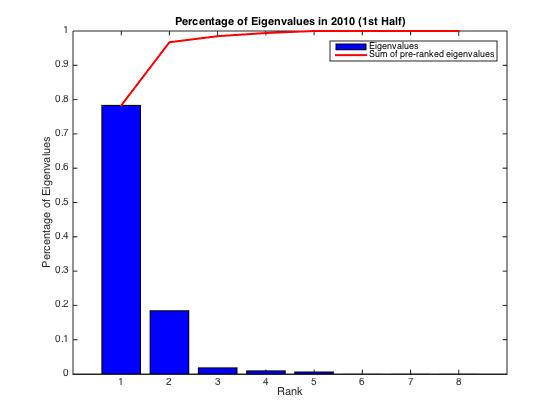


As shown above, the sum of the first 3 eigenvalues is nearly 100% of the sum of all 46 eigenvalues. For this data set, picking top 3 eigenvectors to generate a 3 dimensional data set based on the original data set is a good method for both reducing dimensions and improving efficiency.

However, whether pick 3 dimensions from data bases in all 4 others years is questionable. Thus, running the same PCA process on all other data sets in internal migration is necessary for testing if picking top 3 eigenvectors are always a good performance.





After calculating the eigenvalues in the process of PCA over every data sets, the top 3 eigenvalues are found to cover around 99% of the sum of all eigenvalues. This proves that using the 3 principle components with the top 3 eigenvalues can cover most possible variance of the original 46 dimensional data set.

For this reason, in the process of building all neural networks in this chapter for interviewing data that collected in any year, only 3 principle components are selected and only 3 dimensional data sets are to be trained, validated and tested.

This process largely promotes the efficiency of building neural networks and also can decrease the influence of noise that may exist in components with low eigenvalues.

With 43 dimensions fewer, artificial neural networks can be trained with higher time efficiency and lower vibration.

# 5.4 Modeling

In the last section, PCA is introduced for de-dimensioning interview data into few 3 dimensional data sets. With these data sets, feed forward neural network can be efficiently employed over these data sets.

As for running neural network, this time, the maximum number of nodes for each layers are also set to be 40 and the four algorithms used are:

1. Gradient descent backpropagation (traingd) – Parameter: learning rate (lr).
2. Gradient descent with adaptive learning rate backpropagation (traingda) – Parameters: learning rate (lr), ratio increase/decrease learning rate (lr\_inc, lr\_dec).
3. Gradient descent with momentum backpropagation (traingdm) – Parameters: learning rate (lr), momentum constant (mc).
4. Resilient backpropagation (trainrp) – Parameters: Increment/Decrement to weight change (delt\_inc/delt\_dec).

The Neural Network Toolbox in Matlab is used for building the neural networks.

First, a few tests are run to prove the presumption made above is correct.

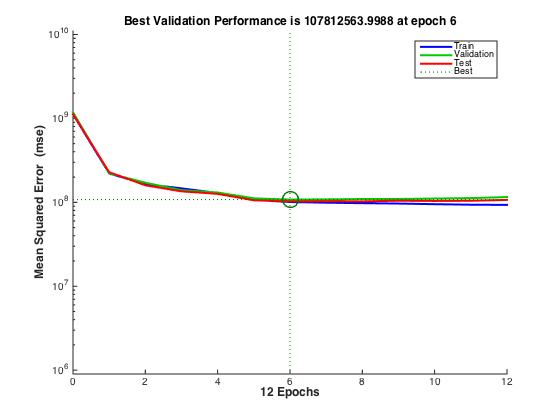
To prove the specified imputation method for individual internal migration data can improve the accuracy of the model, a test is run by building the model based on the data of migrators to Beijing without any reduction on dimensions.

The training process run all possible combination of number of nodes in layers and result in the following data:

The accuracy with the data that imputes 0 for missing data:

|  |  |  |
| --- | --- | --- |
| Mean Squared Error | Number of Nodes in 1st Layer | Number of Nodes in2nd Layer |
|  | 27 | 28 |

With the regression as follow:

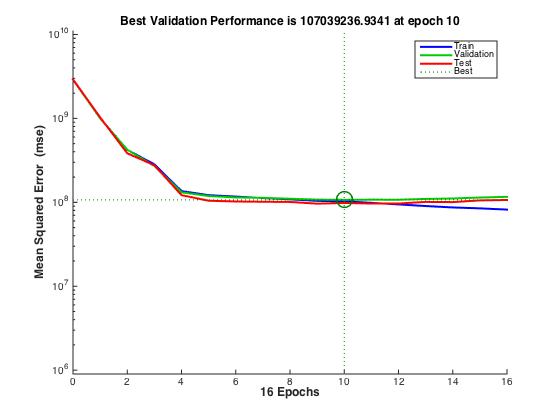


ToBeijing0

The accuracy with the data that imputed following steps in this paper:

|  |  |  |
| --- | --- | --- |
| Mean Squared Error | Number of Nodes in 1st Layer | Number of Nodes in2nd Layer |
|  | 19 | 23 |

With the regression as follow:



Beijing1

As shown above, the steps enhanced in this paper successfully in reducing the error of predictions. Thus, in the following analysis, all data are imputed following the steps described in this paper.

Next, the effect of PCA is to be tested. To compare the performance that without PCA as described above, this section uses the same interviewing sub data set with only data recorded in 2012 of migrators travelling to Beijing.

After Picking the top 3 eigenvalues, the result is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| Propagation | MSE | Layer 1 | Layer 2 |
| Gradient descent backpropagation |  | 3 | 2 |
| Gradient descent with adaptive learning rate backpropagation |  | 2 | 39 |
| Gradient descent with momentum backpropagation |  | 10 | 1 |
| Resilient backpropagation |  | 39 | 16 |

The propagation algorithm with the least mean squared error (MSE) is only about 15% lower than the result trained in the original data set. Thus, the data sets after principle component analysis still keeps most variance within the data sets and shows good performance in comparison with the original data set without PCA.

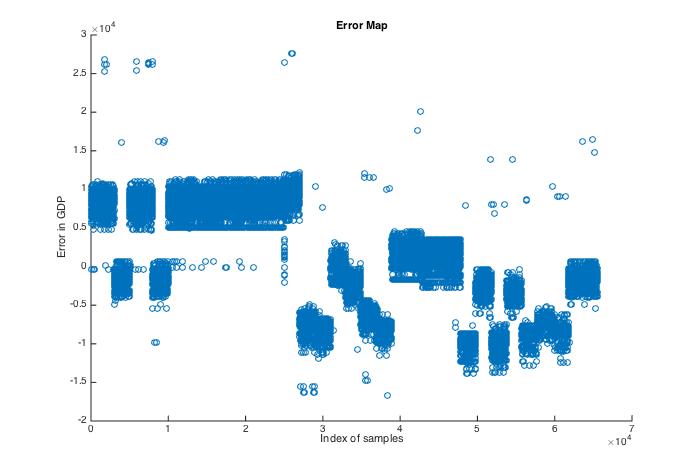
All data sets used later are data sets after running PCA with the selection of top 3 principle components.

With the data methods provided above, artificial neural networks are trained, validated and tested with the interview data in 2012. After iterated over 1600 different topology of neural networks and four propagation methods, the results are shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| Propagation | MSE | Layer 1 | Layer 2 |
| Gradient descent backpropagation |  | 10 | 2 |
| Gradient descent with adaptive learning rate backpropagation |  | 16 | 36 |
| Gradient descent with momentum backpropagation |  | 1 | 9 |
| Resilient backpropagation |  | 39 | 16 |

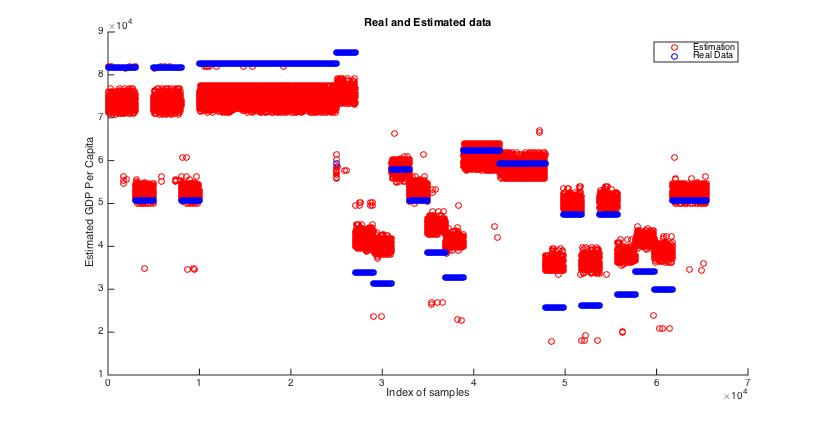
The result shows the square errors are around , which means real errors are around in Chinese Yuan, which can about to decide which province the immigrate about to travel to, since GDP Per Capita in provinces near to each other are mostly larger than .

The scatter map of errors in predictions:



As can be seen from the error map, almost all errors are quite small relative to the GDP per capita in each province.

The scatter map that records the real and estimated data is shown below:



From this img, the scattered prediction reply relatively close to the destination’s GDP per capita.

If the exact destinations of migrators are required for the analysis, the province code of the destinations can he predicted by the distances between origin and each province and the corresponded GDP per capita. The process is listed below:

1. Traversing all provinces and find provinces with GDP per capita with the difference less than 10000 RMB;
2. Select the nearest 3 provinces geographically from the origins of migrators and save them for potential lists.
3. Calculating the possibility of each potential province by corresponded distance with the sum of all distances throughout all 3 provinces.

After this process, 3 provinces with the possibilities can be computed.

# 6 Conclusion and Future Work

This paper focuses on using several methods and data mine tools for the analysis on Chinese internal migration. In the process of analysis, this paper divides the process into two, one in the general view of the internal migration and one in individual view upon it.

One is to use several classification methods for the prediction on general national migration rates between provinces. This process is done by using several data mine methods including random forests, bagging aggregation with c4.5 decision tree, linear regression and feed forward artificial neural networks.

The migration rates in this process contain both inter-provincial and inner-provincial migration, which means the migration rates provided are between the province with all provinces including itself. One point need to be mentioned here is that the word ‘province’ in this paper also includes direct-controlled municipalities and autonomous municipalities but excludes Hong Kong, Macro and Taiwan since the visa or permits are required among these areas.

Another one is for using artificial neural network to analysis the individual intention for the destination of migration. This includes the process of data preparation, principle component analysis and artificial neural networks modeling. Data used for this process must be imputed properly.

For the first process, MIC scores are first calculated for selecting parameters. Since the potential parameters especially in the area of economy and environment in each province are redundant when to be chosen, using MIC scores to deselect parameters that are not correlated to the migration rate shows a nice action in choosing parameters that can make the prediction be away from the vibration caused by over-fitting and only learning from the information that have enough correlation with migration rates.

The built models are successfully in the prediction of migration rates between and inside the provinces in China. As mentioned before, the accuracy can be controlled around 0.04 in measured by RMS and the best method for this process is always random forest for its best prediction accuracy and perfect low vibration among different data sets in this area.

Then, a prediction upon not only one year but few years are introduced. With the same model that built in 2011, even using 2012 data as input, the prediction can be controlled around 0.05 in RMS by random forest. This shows the ability of using one model for prediction of migration rates for future years with the prediction of other economic data.

This paper also tried to use the history data that collected not only in one year before the potential data but also in several past years. However, this does not improve the accuracy as wished. Adversely, the increasing in strong correlated data (parameters in the same filed but collected in different years) decrease the accuracy in some methods, and, as the best method found by this paper, random forests do not show much improvement after be trained and tested by these data. This paper analyzes the reasons should be the potential over-fitted training with too much correlated data and the increasing of noise with the historical data.

In addition, this paper also proposes not to collect too many parameters that within similar fields since this might cause the training algorithm over-fitting on these parameters.

The second process is for the analysis on individual intention on the choosing of destinations for migration. This paper first introduces methods that for the imputation on data sets. Since most missing data (actually, all data that in the data set tested by this paper) are caused not randomly but by the design of the questionnaire, the missing data in certain parameters actually carry meaningful information such like the loss of jobs. For this reason, this paper imputes the data sets by analyzing the reason behind missing data in each parameter. After fulfilling the 23 dimensions of missing data each by its own meanings, this paper also tests the performance with such a method of imputation and without. With the tests on such imputations, a conclusion that this method will apparently increase the accuracy on our further prediction.

As the original representation of provinces, which is a simple list of codes from 11 to 85 hardly shows much difference between the statistics economical circumstances of on province, this paper proposes a method that to use the GDP per capita in representation of each provinces for the destination of migrators. This change shows in contribution in the increasing accuracy of prediction by artificial neural networks.

Since the data sets contains too many dimensions, which will obviously slow own the speed of training neural networks, principle component analysis is employed in these data sets by selecting the 3 principle components with top 3 eigenvalues in PCA. To make sure the major information is reserved, the data generated based on these components is used into building the artificial neural networks with Levenberg-Marquardt propagation along with the same data set but without PCA. Through the comparison between the accuracies, the fact that using PCA in these data sets do not cause the apparent lose of important information for making prediction. Thus, all data sets used later are processed by PCA for the reducing of dimensions.

After the reduction of dimensions, artificial neural networks are used in making predictions towards the GDP per capita of the destinations. After training 40\*40 times for each propagation algorithms and totally 4 algorithms, results with considerable accuracy are generated and shown in prior chapter. The mean error of each prediction is around 10% of the true value of GDP per capita.

Finally, a further method that can predict the exact destination of one migrator rather than the GDP per capita is introduced by this paper by the selection of GDP per capita that near the predicted one and giving the possibility of each by comparing the distances between origin and destinations.

Based on the process before, a general method that for internal migration can be conducted. These process shows very low dependence on priori knowledge and can be used by any similar project with only a few changes in the process of collecting and preparing data. For example, data mine in international migration can also follow the steps introduced in this paper for the prediction of migration rates and individual intention over migration.

Further work can be down in several aspects.

Firstly, the collection of individual migration data does not contain enough data for prediction. The migration rate calculated by these data sets are commonly to be 0 between some provinces. But, actually, with such a huge population in each province, the migration rate can hardly be 0. Increasing the amount of interviews taken will benefit the accuracy of prediction.

Secondly, obviously, almost most candidates took the interviews as recorded in the interview data sets are labors but not students nor citizens in middle class, who are also a large portion in migrator. This problem can directly cost the prediction made in this paper shows more keen to the situation of labors but not the whole group of migrators.

In addition, values of some parameters in the interview data sets can not clearly represent carry the difference between the true information between them. For example, in the question q202 in the year 2012 as mentioned before, the value 2-5 stands for several situations of not having a job, which carries similar meaning (not employed) and only 1 stands for employed. This may easily cause neural networks miss-weighting it.

# Reference

(1) Hu, Xiaojiang, Sarah Cook, and Miguel A. Salazar. "Internal migration and health in China." *The Lancet* 372.9651 (2008): 1717-1719.

(2) [D. Reshef, Y. Reshef], H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, [[M. Mitzenmacher, P. Sabeti]]. Detecting novel associations in large datasets.*Science* **334**, 6062 (2011)

(3) Artificial Intelligence: A Modern Approach (3rd Edition) (11 December 2009) by Stuart Russell, Peter Norvig (p728)

(4) Yearbook, China Statistical. "National Bureau of statistics of China." *China Statistical Yearbook* (2011).

（5）Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.

(6) Harris, John R., and Michael P. Todaro. "Migration, unemployment and development: a two-sector analysis." *The American economic review* (1970): 126-142.

(7)Zhang, Yudong, et al. "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection." *Knowledge-Based Systems* 64 (2014): 22-31.

(8) Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

(9) Breiman, L. Machine Learning (2001) 45: 5. doi:10.1023/A:1010933404324