International Journal of u- and e- Service, Science and Technology
Vol.8, No. 11 (2015), pp.273-282
http://dx.doi.org/10.14257/ijunesst.2015.8.11.27

Research and Prediction of Flowing Factors of the New Generation Migrant Workers in HeiLongjiang Province of China

Qinghe Pan[1] and Xingchi Zhao[2]

[1]*School of Computer and Information Engineering, Harbin University of Commerce, 150028*
[2]*Department of Foreign Languages, Harbin Vocational College of Science and Technology, 150300*
570749130@qq.com, zxc321222@163.com

Abstract

This paper mainly focuses on the analysis and prediction of flowing factors of migrant workers in HeiLongjiang province of China. The feature of the research is that a software system is also developed to assist analysis and research. Sankey diagram and CART (Classification and Regression Trees) algorithm are used as main techniques to analyze data and visualize it. The research methods and software system in this paper can provide support for making decision on labor policy.

Keywords*: Flowing factors, New generation migrant workers, Sankey diagram, CART*

1. Introduction

The flow of migrant workers in China has been the focus of the study, because the flow can always reflect economic problems whether short or long term [1-4]. The workers are mainly composed of a large number of Chinese farmers, so China may pay more attention to migrant workers than other countries. "New generation", the word emerges recently, refers to the farmers who acquire more knowledge, techniques, and educations than older generations. The "new" doesn't mean young. If an older worker owns the above characteristics, then he is "new". The emergence of new generation is not accidental. Firstly, China paid more attention to education on rural people during the last thirty years. When part of these educated people flows into labor market, they have more advantages. Secondly, when the older migrant workers had entered the labor market, they realized the importance of knowledge and techniques. As a consequence, they might be inclined to offer their children a better education. Thirdly, the pressure of competition and the cost of living forced them to keep learning.

New generation migrant workers bring out new problems. Different from traditional methods we need to consider new factors that influence flow of the new generation workers and make prompt analysis and prediction on the main factors that makes them flow.

In this research we will study on analysis and prediction of flow factors of the new generation migrant workers in HeiLongjiang province of China. In view of our survey and more than 3000 questionnaires on migrant workers, it firstly analyzes the basic relations between factors. Based on the basic information and the realistic decision requirements, the decision tree technique is used to further dig information from our data so as to predict values of key flowing factors.

There are three sections in this paper. In the first section it shows the background and descriptions of our questionnaires. In the second section it uses sankey diagram [5] to generate the visualization of data and analyzes the relationship between data. In the third

ISSN: 2005-4246 IJUNESST

section it focuses on decision tree method to make prediction that will deepen the understanding of data and help us make rational decision for the future as well.

2. Background

In order to research the problem mentioned above, a questionnaire survey was carried out. The survey objects were more than 3000 migrant workers. After cleaning data only 1990 effective questionnaires were saved. The content of questionnaire is shown in table 1. All the attributes are considered to be the key factors of giving impetus to flowing.

Table 1. The Content of Questionnaire

	Attribute name	Value range	Description
1	Age	{1…}	The labor's age.
2	Gender	{Male, Female}	The labor's gender.
3	Marital_status	{Unmarried: 1, Married}	The labor's marital status.

4	Education	{
		(Semi)illiteracy,Primary school,
		Middle school,High school(above)
		}
			The labor's education degree.

5	Flow_number	{1,2,3,4,5,6}	How many times have the labor changed the job?
6	Trained_or_not	{Untrained, Trained}	Whether the labor participated in training before work?
7	Organized_or_not	{ Unorganized, Organized}	Looking for a job by self or with others?

8	From_where	{
		In country,
		Outside country and in town (CT),
		Outside town and in province (TP),
		Outside province and in nation (PN),
		Abroad
		}
			Where does the labor come from?

9	Job_category	{
		Planting,Livestock,Forest,Fishing,
		Mining,Architecture,Manufacture,
		Transportation,Retail,
		Restaurant,Service,Other,
		}
			The labor's possible job category

10	Salary	{1…}	The labor's salary.

In Table 1 there are 10 attributes. "Attribute name" column lists the attribute names. "Value range" lists the corresponding possible values of each attribute. And the "Description" column describes the meanings of attributes respectively. In the next section the sankey diagram is used to analyze and display the contents of table.

3. Data Visualization and Preliminary Analysis

Sankey diagram visually reveals the complex relationships between data. Although belonging to displaying technique, it can give us inspiration and prompt for further data mining. So, sankey diagram is used as an auxiliary tool to display and analyze the relationships between data in the research. In order to facilitate the research, we develop a set of software tools. The GUI for sankey part is shown below.

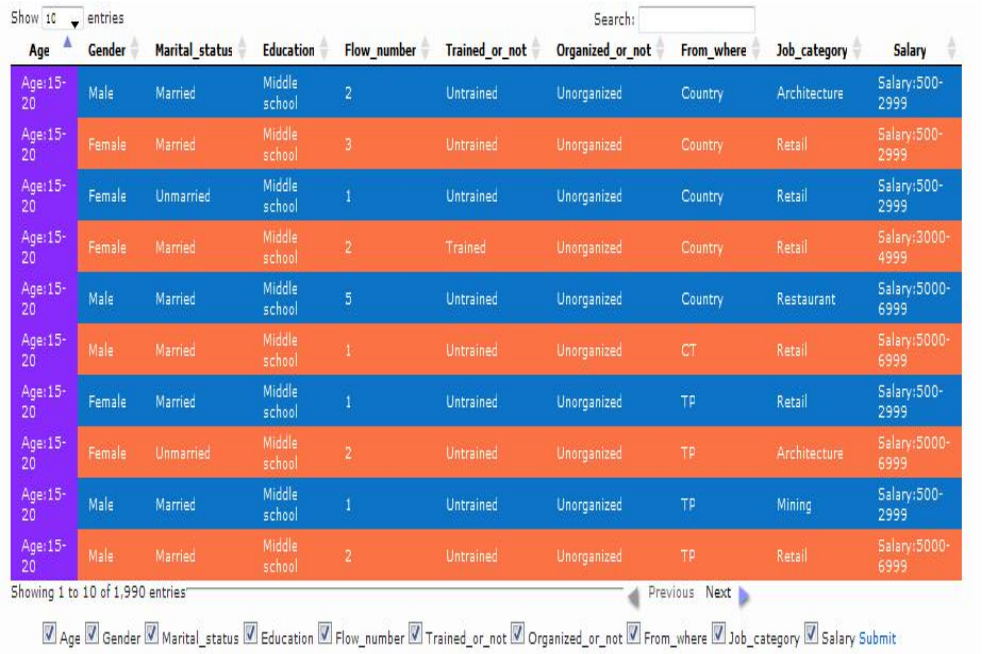

Figure 1. The Interface for Sankey Part of Software System

The interface has two components. The top table shows the basic data according to the format of table 1. Some attributes like "Age" and "Salary" are classified by specified rules, which are based upon the analysis requirements. The checkboxes below two tables correspond to column names. If we want to display the relationships between some columns, just click the corresponding checkboxes and then click "submit". Then the explorer is opened and the sankey diagram will be shown in it. For example, when the "Age", "Education", "Job_category" and "Salary" are chosen and "Submit" (see the Figure 2) is clicked, the Figure 3 will be displayed.

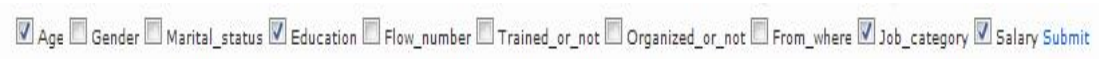

Figure 2. The Situation of "Age", "Education", "Job_category" and "Salary" are Chosen

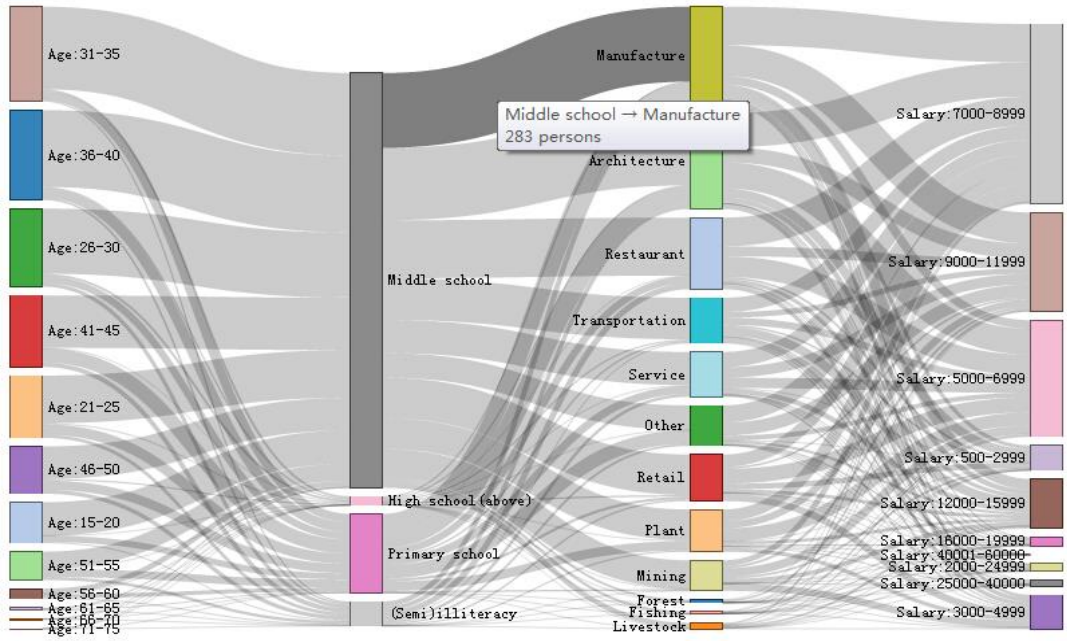

Figure 3. Sankey Diagram of "Age", "Education", "Job_category" and "Salary"

If we click a stripe on the graph, the corresponding tip will be shown to describe the data relation between two endings of the strip. According to Figure 2, the quantity of manufacturing labor on middle school level is 283. The focus of sankey diagram is usually not the quantity but the proportion relationship between data. If the difference between proportions is small or we want to describe relationships precisely, the quantity description will be needed. The following two examples show these situations. In this figure, we can see the labor's education degree is mainly on middle school level; manufacturing labor and architecture labor are similar on quantity; the proportion of labor with salary between 7000 and 9000 is higher than any other salary range and etc.

Choose all checkboxes and submit (see the Figure 4), we will get the Figure 5 showing all data relationships.

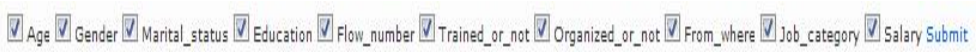

Figure 4. The Situation of All Attributes are Chosen

From the Figure 5 we can identify many factors. The number of proportion will be given to precisely describe the factors.

(1) The majority of migrant labor are young adults. 15-20 years old labor account for 7.74% of the rural migrant labor in the province; 21-30 year old ones account for 26.63%; 31-40 years old account for 35.03%; 41-50 years old account for 22.51%; ages between 51-60 account for 7.34%;60-75 year-old labor account for 0.75%. The age under 40 account for 69.40%, which is the main part of rural labor force.

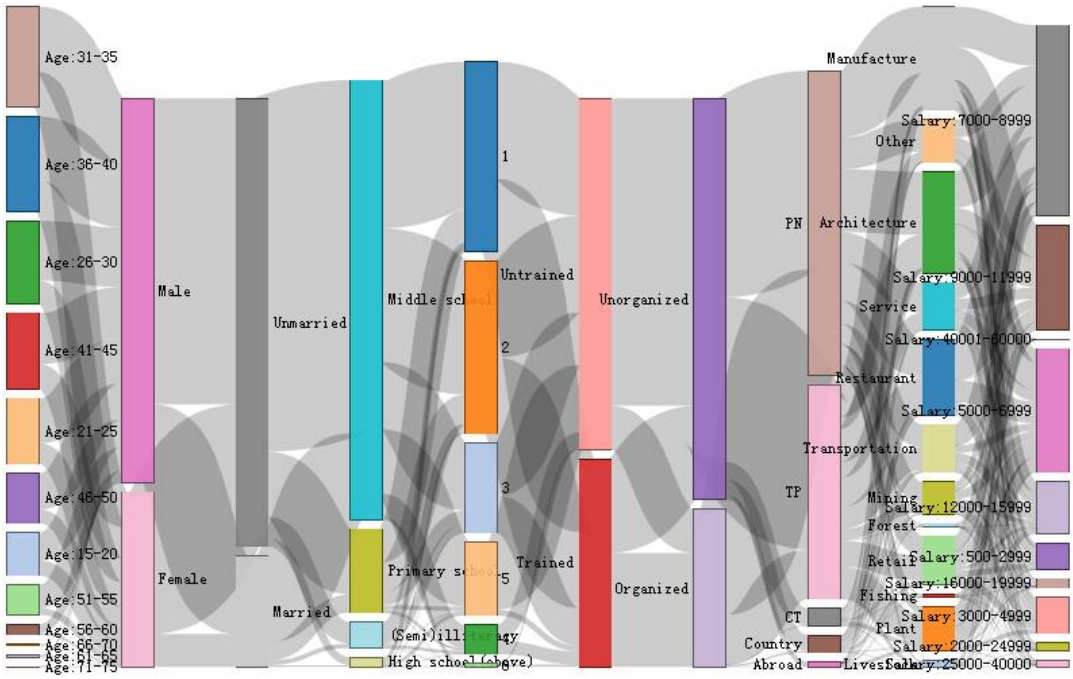

Figure 5. Sankey Diagram of all Attributes

(2) The education degree of migrant labor is mainly on middle school level. (Semi) illiteracy account for 4.89%; primary school account for 15.68%; middle school account for 82.37%; high school(above) account for 1.79%.

(3) The migrant labor mainly concentrate in the province. The labor in country account for 3.11% ;the labor outside country and in town (CT) account for 3.17%;the labor outside town and in province (TP) account for 38.34%;the labor outside province and in nation (PN) account for 54.47%;and the labor abroad account for 0.9% of total employment.

(4) The migrant labor is mainly concentrate in the second and third industry. The labor in the first industry account for 10.30%; the labor in the second industry account for 23.97%; the labor in the third industry account for 58.04%; the labor in other industries account for 7.69%.

(5) There are more men than women migrant labor. The male labor force account for 68.69%; the female labor force account for 31.31%. In our province the amount of rural male practitioners is 2.2 times of the female ones.

4. Make Prediction and Decision by Decision Tree Technique

With the basic information, we can further make analysis and prediction to dig deeper meaning from data. We consider the following three questions on the table below.

(1) For a female labor of age between 21 and 25: How much will she get if she works without training in retail?

(2) For a labor whose education degree is middle school and the range of salary is [7000, 8000]: How old is the labor?

(3) For a married labor: What is the education degree of the labor working in restaurant?

On above questions all the prediction will give the most likely value. We can abstract the three questions with following three patterns:

(1) (Age, Gender, Trained_or_not, Job_category) → Salary?
(2) Age? ← (Education, Salary)
(3) (Marital_status) → Education? ← (Job_category)

In each abstract format above the names come from the names of columns in table. The name with "?" represents the decision object that we want to predict; the names in parentheses represent the known attributes we will assign them values to implement prediction; "→ "or "←" represents the relative position of names in the table and decision direction. Note that all the known attributes are given at the same time and they are unordered. For example in (3) the attribute values of "Marital_status" and "Job_category" will be given at the same time to predict the "Education" value. In (1) and (2) all attribute names in parentheses are unordered.

The above questions can be solved by decision tree technique. In this research the CART [6, 7] (Classification and Regression Trees) is used. The Python [8] is used to implement the algorithm. The corresponding tree will be changed as the decision object changes. So there will be three different decision trees for the three questions above. One advantage of decision tree technique is even only partial attributes are given, the algorithm will still give the relatively accurate prediction result.

We also develop a piece of software to facilitate the analysis and decision process. Its core is the algorithm above. It consists of three parts, including: the interface to choose name of object attribute which will be predicted and other known attributes, the interface to assign hypothetical values to these known attributes, and the drawing engine that uses graphs to draw the decision tree.

Now we can use the software to solve the three questions above.

(1) (Age:21-25, Gender:Female, Trained_or_not :Untrained, Job_category:Retail) → Salary?

The first step is to choose the object decision attribute: salary. The Figure 6 is the interface.

Figure 6. The Interface of the First Step

The second step is to set the value to Age, Female, Untrained, Retail. The Figure 7 is the interface.

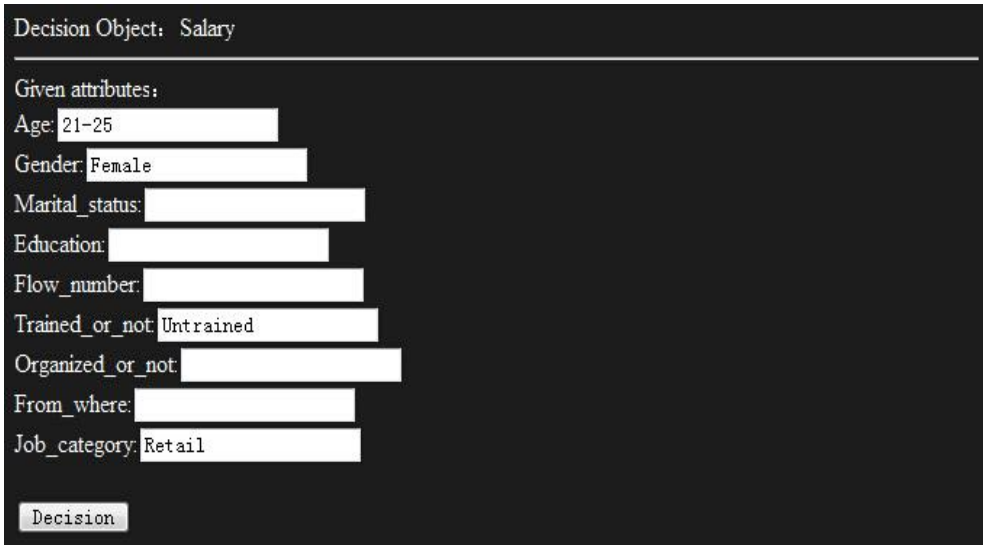

Figure 7. The Interface of the Second Step

By clicking "Decision" button the result "Salary:5000-6999" will be given and a "Show tree" link will display. The Figure 8 shows the result.

Figure 8. The Result of Prediction

Now we get the value of salary so the prediction process has finished.

By clicking "Show tree" link, the decision tree will be displayed on explorer. This is the third step. It's not necessary. But it can give us intuitive clues which indicate the salary value is computed by other known attribute values. The tree is very large and wide so the Figure 9 only shows some parts near the bottom leaves.

The resolution processes of the other two questions are same to the first.

(2) Age? ← (Salary:7000-8999, Education: Middle school)

The answer is 36-40.

(3) (Marital_status: Married) → Education? ← (Job_category: Restaurant)

The answer is middle school.

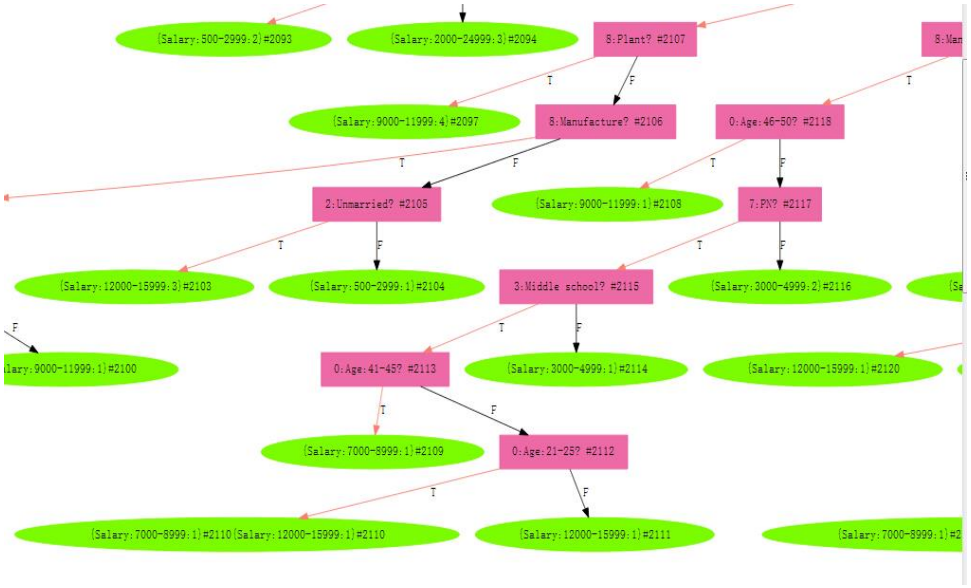

Figure 9. The Decision Tree for Question (1)

5. Conclusion

China has the largest number of labor force in the world. With the rapid economic development, the factors of labor mobility and their impact on China are very important. In recent years, HeiLongjiang, located in northern China, has lagged behind others provinces in economic development. It has more labor outflows than other provinces in China, so it is very important to research the problem of labor transferring in HeiLongjiang. Based on the analysis and research of employee questionnaire data, this study analyzes the factors of labor force transferring and uses the decision tree technique to construct decision supporting system which can provide assistance and reference for the study of labor transferring and its impact on economics.

Acknowledgments

This research is supported by Scientific Research Foundation for Doctor of Harbin University of Commerce (12DL024).

This research is supported by National Natural Science Foundation of China (71240007).

Thanks are due to Professor Zhihao Wang for valuable discussion of this paper.

References

[1] R. Chu, M. Liu and G. J. Shi, "How rural-urban identification influences consumption patterns? Evidence from Chinese migrant workers", Asia Pacific Journal of Marketing and Logistics, vol. 1, no. 27, **(2015)**.

[2] D. H. Liu, H. Y. Li, W. G. Wang and Z. Chuang, "Scenario forecast model of long term trends in rural labor transfer based on evolutionary games", Journal of Evolutionary Economics, vol. 1-22, **(2015)**.

[3] B. Qin, "The Study of Sustainable Labor Migration in China: Based on Lewis Two-Sector Model and a Field Survey", Sustainable Development in Rural China, Springer Berlin Heidelberg, **(2015)**, pp. 55-71.

[4] M. Zhang and R. Rasiah, "Globalization, industrialization and labour markets in China", Journal of the Asia Pacific Economy, vol. 20, no. 1, **(2015)**, pp. 14-41.

[5] P. Riehmann, M. Hanfler and B. Froehlich, "Interactive sankey diagrams", Information Visualization of IEEE Symposium, **(2005)**.

[6] T. Segaran, "Programming Collective Intelligence: Building Smart Web 2.0 Applications", O'Reilly Media, **(2007)**.
[7] J. H. F. Breiman, R. A. Olshen and C. J. Stone, "Classification and regression trees", Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, **(1984)**.
[8] www.python.org.

Authors

Qinghe Pan, Doctor, Lecturer, teacher of School of the Computer and Information Engineering, Harbin University of Commerce. His main research fields include data analysis, electronic commerce, embedded technology.

Xingchi Zhao, Master, lecturer, teacher of Department of Foreign Languages, Harbin Vocational College of Science and Technology. Her main research fields include English education, British and American literature, English for specific purposes and *etc*.