Possible Method for Customer Segmentation in Bath Center Based on Neural Net

Yuxuan Ma

School of Mathematical Sciences Dalian University of Technology Dalian City, China

Abstract: With the coming of data era, even small business has ability to collect and store abundant data, which means doing customer segmentation based on data, instead of experience, is possible. This paper chooses a developing business in Chin—bath center, to make customer segmentation with data. In this research, neural net is used to deal with the distance to fit more with the true psychological distance and create the distance score to get more information from the variables about coming of the customers. Possible formula of distance score is also given as a reference. The distance score turns out to be a better measure of the psychological distance and the neural net has a better graph of the distance data than other methods. This paper gives a possible method for customer segmentation in a completely new business area.

1 INTRODUCTION

With the development of machine learning, more and more businesses use data science to help make profit, and the new subject—business analysis, arises. A plenty of papers focusing on business analysis have been done in the business areas where researchers can easily get a huge amount of data, such as bank and market. Though many small businesses like milk powder [1] and veterinary medicine [2] have the awareness of marketing segmentation, they do not use data or simply use the general method to do it, which may not fit with the situation of specific business area. This paper will take data to do customer segmentation for bath center and give a specific method for the important variables. To help the business make more profit and give better service, specific customer segmentation based on increasing data is necessary. In this paper, decision tree and neural net are used. The reason of using decision tree is that it always has good results in business area. Because of the different standard of each person, using neural net gives a better way to make a measure. By using this method, this paper can turn unused data to sizable profit and improve the customer satisfaction index.

2 METHODS AND STEPS

2.1 Data Preparation

Preparing for data is the most important step for business analysis. A good preparation can both shorten the time algorithm costs and decrease the complexity of the algorithm. Only things given here are general steps and four methods can be used to find errors, which are the most difficult things in data preparation.

2.1.1 Steps

First, use manual analysis mainly. Second, find the mode in data to clean it or define rules based on data. Third, in some specific area, error can be corrected automatically sometimes, but usually need people interference [3].

2.1.2 Methods

(a) artificial method

(b) specific algorithm

(c) solution to specific area

(d) data clean in area not specific

2.2 Feature Selection

Before doing the actual operation, it is necessary to figure out the concept of feature selection for the reason that different people have different definitions on it. There are many similar concepts and sometimes scholars take them as the same, such as variable selection and feature extraction. For example, in the paper of Delta Whistler Resort, variable selection refers to the problem of selecting input variables that are most predictive of a given outcome, and feature selection refers to the selection of an optimum subset of features derived from these input variables [4]. This paper adopts what most scholars agree on as Su Yingxue did and defines feature selection as selecting the most relative subset from the input set [5].

angela@cas-harbour.org

Possibly useful method will be given and how to select the method is based on the data. The advice is to try most of them and observe the result to see which is better. Two major steps of feature selection are generation procedure and evaluation function which are given below.

Generation Procedures:

Complete
Heuristic
Random
Evaluation Functions:

Distance Measures
Information Measures
Dependence Measures
Consistency Measures
Classifier Error Rate Measures [6].

Combining the generation procedures and evaluation functions can get some specific methods.

2.3 Building Distance Score through Neural Net

Absolutely, bath center belongs to the modern service industry. More convenient transportation always brings more profits. According to the consumption motivation of customers, customers come to bath center for convenience or enjoyment in common and different demands lead to different acceptable distance. For the reason given above, the author assumes that the data about distance is in the subset which is selected after feature selection.

In some similar situations, some scholars just use distance as a variable to do the customer segmentation. Actually, distance is to find out how much money and time they spend getting to the bath center. However, it is impossible to value these two factors by just using the distance in a straight line, especially in Dalian like cities. Dalian is a hilly peninsula, which makes many roads there at different height levels. In this case, people usually cannot get to the destination straightly. For example, if someone wants to go to the market across the Lingshui road, he must get to crossroads at his right hand and then walk up along the slope. As a result, taking city traffic network into account is essential and some details will be given below.

How to get the data of traffic is the first problem to be solved. It is difficult and unnecessary to get real-time data of traffic all day long. Instead, a common or ideal state of the traffic network and some influential factors are used to calculate a distance score to evaluate the real 'distance' in their mind.

Obviously, how customers come and where they live are needed. Questionnaire is an effective way to get the data. Another method recommended is making every customer a member with free membership cards and differentiating the level of the card to differentiate the customers. In this case, it is much easier to get the information of the customers. Knowing the address and method of transport, estimating the cost are not hard. Then assume data which is like the table below to help explain.

TABLE I.   ORIGINAL DATA

Customers ID	Visit time	Service	Price	Address	Method of transportation	Time cost (min)	Money cost(¥)	…
11234	2019.2.13 18:06	Bath	68	No.1 Street	Driving	10	10	…
11234	2019.2.13 18:06	Massage	128	No.1 Street	Driving	10	10	…
11234	2019.2.13 18:06	Tea	88	No.1 Street	Driving	10	10	…
11234	2019.2.28 20:47	Bath	68	No.1 Street	Bus	30	1	…
11234	2019.2.28 20:47	Tea	88	No.1 Street	Bus	30	1	…
32436	2019.5.26 22:01	Bath	68	Xi'an Road	Bike	2	0.3	…
…	…	…	…	…	…	…	…	…

Data mostly looks like consumption record. To use it, understanding the data and grouping them in different level are indispensable. Each customer has many visits, which means coming to consume, and each consumption has many services. Each visit has a method of transportation and what should be done is counting the number of times of each method of transportation and ranking times for each customer. Take the first as the most frequently-use and the second as the spare. Both of these two transportation have a distance score. Two simple formula can be achieved by using money and time cost.

$$D1(\text{distance score one}) = (M1(\text{money cost one}) +1)*T1(\text{time cost one})$$

$$D2(\text{distance score two}) = (M1(\text{money cost two}) +2)*T1(\text{time cost two})$$

The reason that using one plus money cost instead of only money cost is to avoid zero cost making distance score zero. Both the time cost and money cost are estimated under ideal condition and then the author takes the influential factors in consideration and uses them to create the weight of two distance scores.

$$D (\text{final distance score}) =Wt1* W11*W12 *…*W1n*D1+Wt2*W22*W23*…*W2n*D2$$

W1k=Nk*w1k

W2k=Nk*w2k

(k represents anyone of the integer smaller than n)

N1+N2+…+Nk=1

W1k and W2k are the pair and w1k and w2k are also pair. Each pair of 'w' represents an original weight getting from one factor. Basically, factors must contain the times they use the transportation (Wt1 and Wt2). Commonly, factors happened regularly or last long time like traffic jam and road closing down can be also used to create the weight in a new aspect. If data is adequate, it is also possible to add some changeable data, such as car accident and weather. Because different factors have different frequencies of occurrence, each dimension of weight should be different. To achieve this, Nk, which is a score of frequency, is used to build the new weight－'W'. The way of using data to attain the final weight will be showed below. Here is a simple example without considering Nk. Instead, original weight is used to make it clear and easy to understand.

D=Wt1*0.5*D1+Wt2*0.5*D2

If the weight, 0.5 and 0.5, is about if there is a traffic jam near the bath center and his first choice of transportation is driving and the second is cycling. Then there is a data saying that one day traffic jam happens and he rides to the bath center and what should be done is to increase the weight of the second and decrease that of the first.

To achieve that goal and get suitable weight, using neural net is a good choice.

A neural network is a collection of linear threshold units that can be trained to distinguish objects of different classes [7]. A neural net has many hidden layers, one input layer and an output layer. It continuously adjusts the weight to achieve the last goal. A simple graph of neural net is showed below. To get more information and details about neural net, read [8]is a good choice.
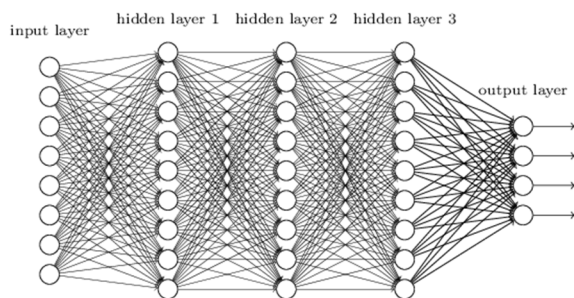

Figure 1 Neural Net

Here the distance score needed to be divided in two stages. First, use the neural net to adjust Nk. And then use another neural net to adjust w1k and w2k. They should be put into two different neural nets and be combined at last. Each stage has similar steps. Continuously put the data in the algorithm and go through it and the answer can be gotten. Certainly, it is possible to use more functions or algorithm to deal with each weight to get a more accurate result. Finally, combine the weight by using simple function and distance score can be used as a variable.

2.4 Decision tree

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules [7]. Decision tree is the most frequently-used method in business analysis, so here the author uses the decision tree to make the customer segmentation. Specific algorithm will not be showed and more details can be got in [9].

Finally, customer segmentation is completed. Business can use this to see what kind of customer they have and many details of each kind of customer. By using the customer segmentation, business can make the strategy directly.

3 DISCUSSION

The method above is customized for bath center. According to the data and the experience of businessmen, the method emphasizes the importance of distance－using distance score to make a better measure for these important variables. As a result, the method performs well in this specific area. This method can do well in other areas where distance is vital by just changing some parameters, which means the method can be used as a general way to deal with the distance in many different areas. For example, this method can help convenience store do site selection. The business can distribute questionnaire to collect data and use this method as an aid. Though for some companies, such as banks, distance may not be such an important factor, this method can also be used to optimize the customer segmentation by making a more suitable distance measure.

4 CONCLUSION

This paper discusses the customer segmentation in a completely new business which is called bath center and gives an outline of the method. In addition, neural net has been used to optimize the distance between bath center and the home of customers and create the distance score based on the original data. However, this paper did not contain the specific algorithm and functions and just used the simple method of data mining and decision tree to realize the last classification. Using decision tree may cause many problems such as over-fitting. To improve the method that this paper has talked about, still many things can be done. Using more complicated data mining methods or simply improving the functions of decision tree are both efficient ways to get a better answer. Top 10 algorithms like SVM and ANN are good choice to attempt [10]. Using special methods to deal with the other variables to make an even much better segmentation is also a good try. For example, also use neural net on other variables. Many problems are still waiting to be solved.

ACKNOWLEDGMENT

REFERENCES

1. Bright Prospect of Food Marketing Segmentation[J]. Chinese Businessman, 1999(2):75-75. (in Chinese)

2. H. Y. Bai, Professional Marketing Segmentation of General Trend of Veterinary Medicine[J]. Chinese Abstracts of Animal Husbandry and Veterinary Medicine, 2010(3):57-58. (in Chinese)

3. F. X. Yang, Y. C. Liu, Z. H. Duan, An Overview of Data Cleaning[J]. Application Research of Computers, 2002(3). (in Chinese)

4. I. Guyon, NIPS 2001 workshop on Variable and Feature Selection: [http://www.clopinet.com/isabelle/Projects/NIPS2001], Dec. 6-8,2001.

5. Y. X. Su, Research on Feature Selection Methods[D]. Graduate School of National University of Defense Technology, 2006. (in Chinese)

6. M. Dash, H. Liu, Feature selection for classification[J]. Intelligent Data Analysis, 1997, 1(1-4):131-156.

7. G. Thamaraiselvi, A. Kaliammal, Data Mining: Concepts and Techniques [J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2006, 5(4):1 - 18.

8. M. T.Hagan, M. Beale, Neural network design[M]// Neural Network Design. 2002.

9. J. R. Quinlan, Induction on decision tree[J]. Machine Learning, 1986, 1(1):81-106.

10. X. Wu, V. Kumar, J. R. Quinlan, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1):1-37.