Research on method of the data classification with an example

Yue Yang

Eberly college of science, Penn state University, state college, United State

yvy5412@psu.edu

Abstract. The huge data from real life is a problem amount the researcher. So, it is important to let us divide them in out project. This article wants to talk about the way of data classification. The article using different method to shows how these methods works, start from the simple word dividing and a little complex method which using the root to be the key word. Then the article shows the complex PCA or k-mean dividing which is always appear with the mechanize learning. Then the article using a true example to shows, how this different way of dividing the data can work for different situation. This article can truly increase the speed of the researcher 'work and increasing the accuracy of the data analysis by letting the data to be more methodical.

Keywords: Data Dividing, Data Analysis, PCA.

1. Introduction

Data are important. Many company or research center collect lots of the data from their life. However, because of the huge amount of the data, the data classification become very important. Miaomiao Lou just point out that the important of the data classification with the medical field [1]. In fact, the data classification become the key point of the scholar to study. The research Yun Shi talks about the data classification base on rough set [2]. Liu and Jiang also talk about the data dividing among the geography field [3].

All the things should use the data to be the foundation. Data is the base things for the not only the company but also the government to deal with their work. However, the raw data might be not good. The raw data is always very large and contain many of the way to deal with that. The data might have thousands of rows and column. It might contain a huge number of the variable and data set. So, there are several ways to deal with that, the first one which people usually use is the things called the data cleaning. The data cleaning is the very important process which use in anywhere. The data cleaning is the process in everywhere. However, the thing this paper uses and this paper wants to talk about here is not the data cleaning. It is the stratified of the data. Many data have many variables. Each variable has different level. After the data cleaning, if the company want to see the data more cleaning and clearly, the segment of the data is necessary. Some of the data have many variables, these variables contain a lot information and easy to miss. So, this paper can use them to help us. This paper can divide the data into several group depend on the data itself. Many variables have different sides. So, if anyone need to use the data for us, they can use that to be a useful tool.

2. Description of theory

2.1. Simple way of data classification

To make the data into several categories, one of the easiest ways of that is using the different variable. Different variable has different level of the data, not only the Nominal variable, but also the Numeric variable. For the nominal variable, it is a little bit easy to recognize. Each word of the data can represent one level of the data. So, this paper can easily divide the raw data into several part of the data category. For the Numeric, the things can be a little bit change. It is impossible for us to divide the data into several part directly. This paper can first do a little bit classification at the beginning. It can divide the number into several group depend on the data here use. This process should include the real world meaning of these data and the actual availability of them. This paper needs to consider the first one which is the real world meaning because here do not want to make useless group. For example, if the 50 score and the 65 score have the different meaning which the first 50 score means the students do not pass the exam and the 65 score mean the students pass the exam and got a C, it is not a good idea for the people to divide these two students into a same group. Also, if the students who all pass the exam and their score is 66 and 65 which mean they both not only pass the exam in this time and get an C in the grade book, it is not a good idea for us to divide these two students into different part. Also, for the second idea, the availability, this paper needs to use the data in a correct way. This paper does not want to divide the raw data into 300 group, in fact this category has no actual difference between the new data frame and the raw data frame. The researcher or the company manager cannot get the information which they want to use for their real life from this still complicated data. So, this paper needs to divide the group in a correct and useful way which means the number of the group should not so much. Of course, the information might lose in the decrease of the number of the group when people are dividing, but when this paper can combine the real world meaning of these data and the truth of the data frame category, people can get a little bit balance between these two parts in my data classification. Similarly, it can be seen Hongzhi Wang also talk about the process of the simple data sorting [4].

2.2. A new easy way of classification

These two things are the most important things for the data dividing. So, this paper need to talk about the basic things about the data dividing. The simplest way of the data dividing is to use the data itself. But sometimes this paper needs to use another way to improve out data frame. Most of the variables are very clear and easy to read. However, sometimes the data description of the variable might be very complex. For example, the description about the product can be one of the interesting variables, but it contains a whole sentence to descries. It is nearly impossible for us to use the whole sentence to be the key for us to divide the data into several groups, because each sentence is hardly similar and it is nearly impossible for us to find the two sentences with the actually same one. So, this paper needs to do somethings with that. One of the useful ways is to exact the keyword or key things from the sentence. For example, if the data frame is the data from the food which are sealing in the market and the description is about the detail of this product, it is a good way for us to check the key words like the vegetable, meat, fish and somethings like that to help us divide the product into several group from the huge number of the raw data and raw data’s variable. Also, it is not easy for us to deal with this. This paper can also use the root of the word. It is an interesting thing because mostly the root represents the meaning of the word. It is even better than using the word instead because people can include many similar words in one root. This thing can help us to divide the data in a faster and clearer way. However, this paper needs to face the problem base on this new way. Although the root includes much more information and words, it can contain much more useless information in the data set. For example, if this paper wants to talk about the type of the product also are sealed in the food market, it is a useless for us to divide the colour to be one of the keywords. So, this paper also needs to clean a little bit of the data after this paper uses it. This paper needs to remove the useless information. A very useful way to deal with such useless information is by find how many times they occur in the data frame. For example, this paper can ignore the data with under 10-time appearance because this paper can treat them to be the less important part of the data set. This number of the new divide can be a little bit easy to change if this paper has different volume of the data set. The larger data frame this paper has, the bigger value the people use. Similarly, Xu talk about similar way with the machine learning [5]. while Solorio-Fernández talk about the similar process [6].

2.3. actual sorting operation

The true way to do the data sorting is a little bit different, this paper has seen similarly way in many other researcher articles. Aggarwal talks about in the book "data cleaning", the train and the test set in useful [7]. Xie also talks about the important of the cluster [8].

Then this paper uses the k-mean to deal with that. After this paper gets a matrix contain the different value of the data set, this paper can get the use of the k-mean. Of course, before this paper uses the k-mean this paper needs to use the matrix of the data set. This matrix should be the combination of the last use of the key word or whatever this paper uses for divide the data set. Because this paper is using the K-mean, so this paper needs to determine the model's n. It is a long process for us to choose the n because this paper needs to run the code of the model by each n in that time. Each time when run the model's code this paper can get a number, this paper represents of this model's average silhouette score. This paper can use a lot of number here but this paper always starts from n=3 to n= 10. Here the paper can see how that work in our example this paper will talk about in the next section. When this paper back to the k-mean, here might also meet some of the latent problem. When this paper chooses the best k-mean n from the average silhouette score, this paper need to still look at the number of each cluster. It is possible that when this paper chooses several of the highest average silhouette score n but then this paper finds that there is some of the problem exist here. This problem can be the very small amount of one of the clusters. It shows that this process of divide might is problematic. If this paper has 3 groups with 1000 number of the data and this paper have another group with only 10 amounts of the data, it seemed quite problematic. Opposite situation might happen in here, too. If this paper has 1 group with 1000 number of the data and this paper has another 3 groups with only 10 amounts of the data, it also seemed quite problematic. So, this paper can edit a little bit after this paper look at the real number of each cluster and get a best one. After this paper chooses the n, this paper can run the model of that. this paper can use PCA to make the data better in the scene. Also, this paper can use the data visitation to help us in the last part.

3. Example of the truth

3.1. Dividing of product

Daniel (September 2017) wants to use a new variable to make data frame more clearly. So, this paper can see it in a better way. In the original data set. The product uses the Stockcode to be the only one thing to identified. Also, the variable "description" can be an extra thing for the talking and introduce it. The author wants to use description to help him here.

The variable "description" is the combination of some words. So, to make the data clearly, this paper needs to extract words from this sentence. The author first extracts the names from the data set. Both the proper and common one. Then the author takes out the root from the word gather the set of names have the relationship with this particular root. After that, the author needs to count how many times this root appears in the data set. On the other hand, if the model finds there are many words list on a same root, the author suggests the shortest name is the keyword of this one. Simliarly, the model will choose the singular when the singular and plural variants of the word appear at the same time. That is the idea of deal with the word description. So, the author begins to work. First step is to check the list of the product. The author gets three things. They are "keywords", "keywords_roots", "count_keywords". The keywords are the list of extracted keywords. The keywords_roots is the dictionary; the keys are the roots of these keywords and the values are words which have relationship with those roots the last one count_keywords is how many times the dictionary list here. This graph shows the count_keywords. Here is a list of several important keyword the author takes out from the description. However, this paper can see there are a lot of key words here. In fact, there are over 1600 keywords here and the most common keywords appear in about 200 words. Many of the keywords here is useless. For example, the keyword “colour” cannot help us here. So, the author decide to use the keyword only if this keyword appear more than 13 times in the data set. this paper can see from the code that the author removes the stop-word from the NLTK package. Then remove some of the very common word in the list like the many kinds of the colour. Then the author gets the list after deletion.

After that the author creates a matrix, use 1 to represent this word occur in this data and 0 to represent it do not appear. After the author creates the matrix, he finds that to take the price level into the matrix can let the matrix become more clear and acute.

After making the matrix, the author want to use the create clusters of products. He divide the different item into different groups. Because the author uses the binary encoding, cosine similarity/distance is the useful and common way to measure the distance. Then the author just gets the k-mean matrix. The author first uses the n equal to 6. However, once the author finds that using 5 will get a better result. Then he change to 5 at last. This is possible because each time the result is not complete same when this paper compares with before. The result of that is complete random.

After the cluster is made, the author shows the different plot of that. The cluster divide the data into different group. In fact, the group diverse clear in some situation. First the author shows the silhouette scores of each element of the different clusters. Then the author shows the number of appearances in the different cluster.

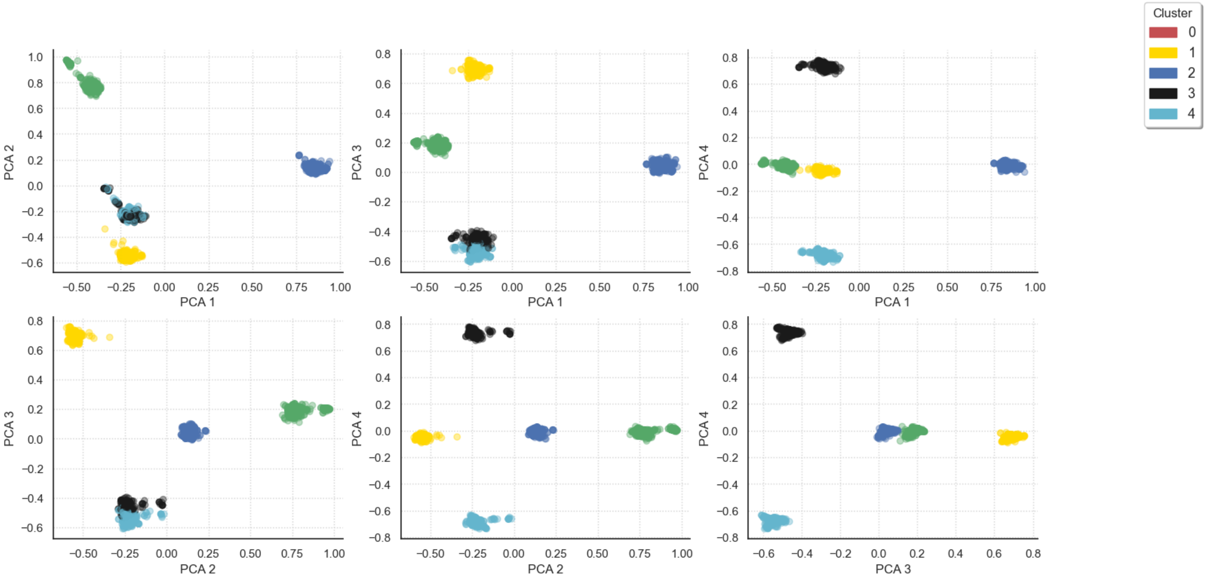

Figure 1. PCA for product.

When use the PCA which is this figure 1 shows, the author finds there need over 100 components to explain about 90% of the data variance. So, the PCA is not work very well here.

After this section, the author completes make the product into different group. So, the next section, the author starts to make the customer into different section.

3.2. Customer dividing

After dealing with the product, this paper needs to look at the different customers. There are a lot of data in there so this paper needs to divide them into groups. First, the author need to input product cluster into the segment. The author create a new categorical variable categ_product to represent it. this paper can see the last column is just this new variable.

After that, the author create a new variable which is called the categ_N. It contains the amount spent in each product category. Because in the last section this paper can see that the author use n equal to 5 in such section, so here the N is also 5 which means there are 5 more columns appear here. The model result shows, the category 5 have the highest revenue and the category 1 have the lowest revenues. Other three category have similar revenues. The author successful divide the information relate to an induvial purchase into one line, but now the author want to move on. He decide to collect the information with the special order and put them into one entry. So, the author create a new data frame which include the amount of the basket and how they distribute over all the 5 categories. So, this paper can get a new data frame with the id, invoice number, price, date and the 5 categories.

The author wants to use the 12-month data to help he build a model to predict the customer's activity. He uses the first 9 month to be train the model and use the last three month to be test the model. In the second step, the author combine the data with one customer together. So here he can know the basic things about this customer, like the date, the maximum or the minimum cost, the average, or the total cost. The author also adds the price of the cancel order and how much it takes in the total order. Finally, the author adds 2 new variable shows the day after the first purchase and the last purchase.

The result shows 95% customer cancel 11% of the item, they buy less than 241 product and visit less than 9 times. Some of the plot shows part of this result.

The author also finds there is a special customer which only buy once. These kinds of customer take about half place in the market. So, when he does the segmentation, he will put these buyers into their segment.

Now the author starts to do the segment of the customer. He uses the data frame "transactions_per_user" here to be the summary of that. So, then the author starts to make the cluster of the customer. The result of the agglomerative clustering is not good, the author decides to use k-mean instead. During the process, this paper can see that the k=4 get the best result.

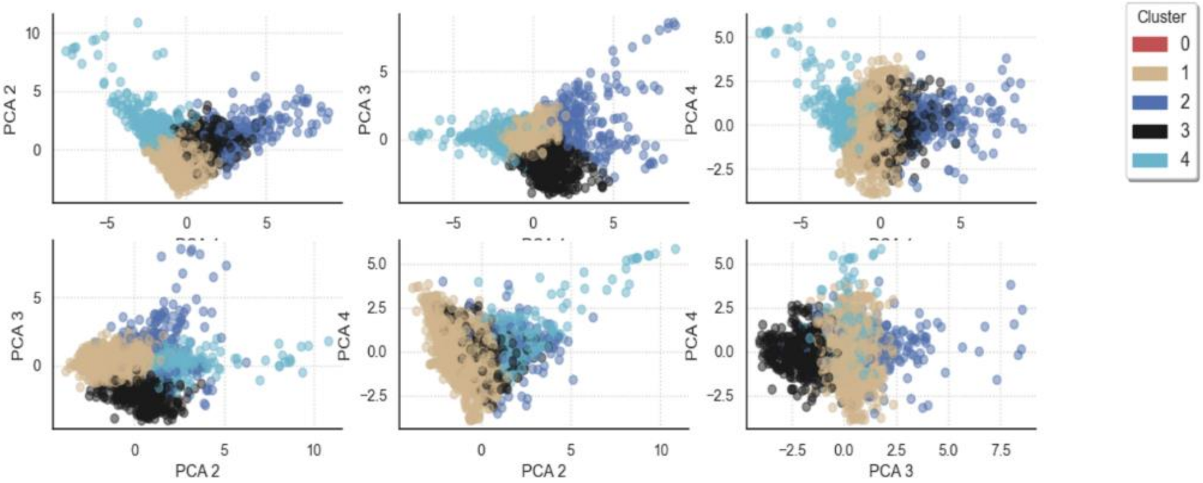

Figure 2. PCA of customers.

First, the author shows the PCA result which this paper can see in figure 2. These clusters look nice and separate into different place.

Such data visualization is easy for us to get a clearly result. So, this paper can see different colours point which represent different group have clear different and boundary between them.

Second, the author starts to look at the silhouette scores to see the quality of the separation.

After these two steps, the author proves that the different customer in different groups is disjoint. But he still needs to understand the behaviour of the customer. Then the author averages the contents for the data frame by first choose the clients in the different group. At last, the author re-organizes the data frame from different cluster.

Finally, this paper can have the plot. this paper can see many of information. For example, the segment 1 mostly spend on the cate2,3,4,5. Segment 4 also spend on 2,3,4,5.

Then the author does the RFM analysis of the data frame here. It shows the cluster 2 have the higher frequency and recency. At last, this paper can get a whole lot of these customer's segment.

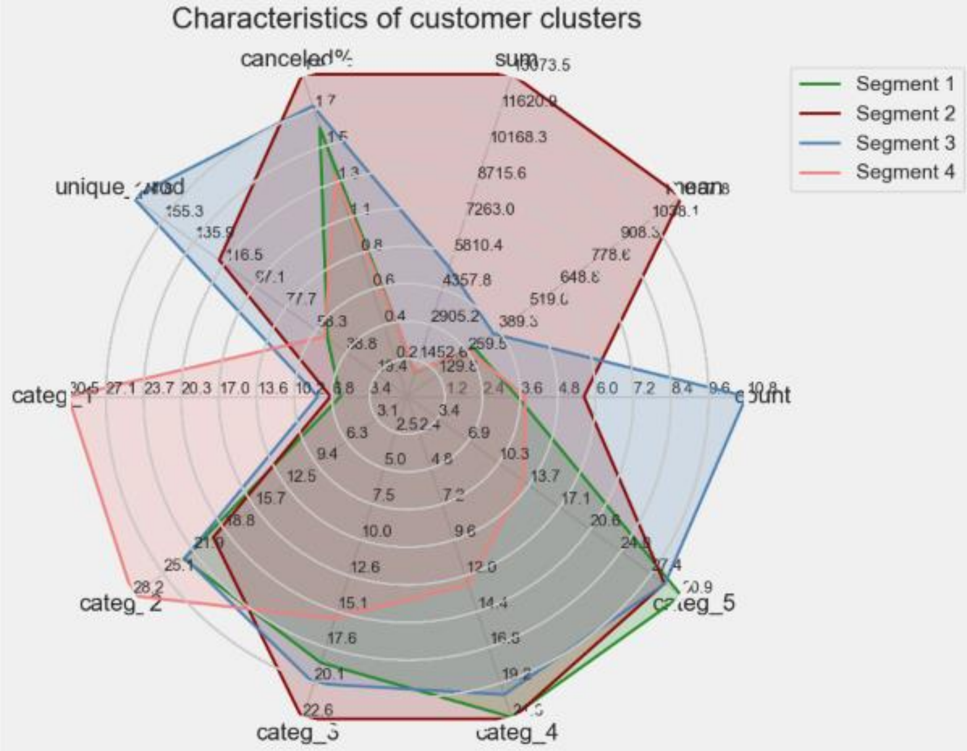

Figure 3. final plot of custom clusters.

This paper can summary the data cluster with the Figure 3.

Segment A has low spending with many kinds of product.

Segment B spends a lot in the category 3,4,5.

Segment C has low spending with category1,2.

Segment D has lots of spending in the category 2,3,4,5.

So, this are simply useful part in anywhere. Also, this part is most like Guo using the method in his article or like Zhang [9] using with the smart grid field [10].

4. Conclusion

This paper talks about the data classification among the true life with their data. Like Guo using the method in his article or like the Zhang [9] using with the smart grid field [10]. Showing some of the process or way to help people getting start of the data classification. First, this paper starts to look at the basic way of dividing the data by variable itself, both for numerical and categorical. Then this paper starts to look at the way to dividing the data by some easy way like searching the key word or even the key root. After that this paper starts to make the real example of how to really use this way to divide a dataset. This research gives people a very useful way to help them dividing the data before they run any type of the model, because it always has their utility. For the more complicated dataset, the easy way of data dividing in the second part of the research can also help researcher a lot.

However, the passage only talks about a small amount of way of the data classification, and it is the future work of me.

References

[1] Lou Miaomiao, Liu Danhong, Wang Xia, Yang Peng, Tan Zhijun, & Liang Ying, etc. 2013. Classification and description of basic data of health statistics. *Chinese Journal of Health Information Management* (1), 5. pp26-30

[2] Shi Yun, net, Sun Yufang, & Zuo Chun. 2000. Spatial Data Classification Based on Rough Set. *Journal of Software*, 11(5), 6.pp84-87

[3] Liu Ruomei, & Jiang Jingtong. 2004. Research on the principles and methods of geographic information classification——taking basic geographic information data classification as an example. *Surveying and Mapping Science*, 29(B12), 4. pp84-87

[4] Wang Hongzhi, Sun Ming, Qi Zhixin, & Gao Hong. 2018. Data classification method, device and storage medium. CN108564137A.

[5] Xu Chenguan. 2012. Research on the classification of Chinese universities based on data mining clustering technology, *Huazhong University of Science and Technology*. pp13-19

[6] Solorio-Fernández, S., Carrasco-Ochoa, J.A. & Martínez-Trinidad, J.F. 2020 A review of unsupervised feature selection methods. *Artif Intell Rev* 53, pp907–948.

[7] Aggarwal, C. 2015. Data Classification. Data Mining. *Springer*, Cham. pp1-25

[8] Xie, J., Girshick, R. &amp; Farhadi, A. 2016. Unsupervised Deep Embedding for Clustering Analysis. Proceedings of the 33rd International Conference on Machine Learning, *in Proceedings of Machine Learning Research* 48:478-487

[9] Guo Dapeng. 2010. Application of genetic programming algorithm in data classification.

[10] Zhang Xiangquan, Li Yunge, & Tan Weidong. 2009. Data classification and network structure related to primary equipment in smart grid. *Grid and Clean Energy* (10), 5. pp29-33