# Users Classification on Broadcast and Television System Based on Statistical Analysis System Software

Binbin, Wan, Taozheng Zhang, Jianping, Chai, Min, Zhao, Zhenlong, Zhang
Science and Engineering Dept.
Communication University of China,
Beijing, China
wwbenben@126.com, zhangtaozheng@cuc.edu.cn, jp_chai@cuc.edu.cn, 403271863@qq.com, 1072547168@qq.com

***ABSTRACT:*** *In this paper, we take the guide data and the program data from the users of digital cable television programs as the experimental data to carry on our experiment. And we use the SAS software platform which is very efficient in data analyzing to realize the classification of our user to different parts. So we can achieve our destination of personal recommendation and precision advertizing.*

**Keywords:** SAS; clustering algorithm, user classification, broadcast and television system, data minning

**Received:**



## 1. Introduction

The rise of the data mining theory bring a huge change to marketing approach of all walks of life .They converted the original methods which focus their thoughts on the revenue from optimizing production or sales to Customer Relationship Management(CRM) and precise marketing. In recent years, personalized recommender system and precise marketing system pushing many company forward. According to people's consuming behaviors to find out their latent interests so that to recommend the most possible things that they liked. Based on the above theory, we can provide the consumers a quick and convenient consumption experience. To realize the personalized recommender and precise marketing, it is the user classification which is based on user's interest that is needed. There have been some achievements around the world in the user classification area. Wei Liuxi uses the SPRINT which is a branch of the decision tree to launch the information digging from the big data to realize the user classification [1]. Zhang Heng based on the consumption KPI to classify the telecom users to different parts which can provide the target user to the precise marketing. Slanjankic use SAS and data mining to realize the chart display and cluster result analysis [2]. Data mining technology is widely used in many areas such as banking business and E-commerce industry but the radio and TV industry [3]- [4]. So, this paper based on the digital TV viewership data and the program data to launch the experiment. The platform is SAS software and the arithmetic is K-means cluster. The realization of the user classification can be a support of the decision support system of radio and TV industry.

## 2. Background And Related Work

Neuman (1997) lists content analysis as a key non-reactive research methodology (i.e.non-intrusive) and describes it as: "*A technique for gathering and analysing the content of text. The 'content' refers to words, meanings, pictures, symbols, ideas,*

*themes, or any message that can be communicated. The 'text' is anything written, visual, or spoken that serves as a medium for communication*" Mika Rautiainen and his colleagues introduced data mining system and broadcasting services for Finnish DBV broadcast stream national channels. Their system allows the access for media broadcast fragments as picture quotes via generated word lists and provide content based recommendations.

Content analysis is used to study a broad range of 'texts' from transcripts of interviews and discussions in clinical and social research to the narrative and form of films, TV programs and the editorial and advertising content of newspapers and magazines.

## 3. Experiment

### 3.1 Introduction of the arithmetic and software

#### 3.1.1 Clustering algorithm

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file for "*MSW_USltr_format*".

Cluster is a processing that dividing a data set into different groups which are similar internal but are different external. It is a classical algorithm which is widely used [5]. The following picture is a display of the process of cluster.

Clustering algorithm has many branches. Considering the big data set and the result to be, we chose the K-means algorithm. Its process is as the following:

a) Defining a original center for every category, so, it has $k$ centers;

b) Allocating the samples to the nearest category by the Minimum distance criterion;

c) Using the mean value of the category as the new center;

d) Repeating step (2) and (3) until there are no change about the center;

e) By the end, there have $k$ category.

The formulas involved are as following:

Allocating the samples to the nearest category by the Minimum distance criterion:

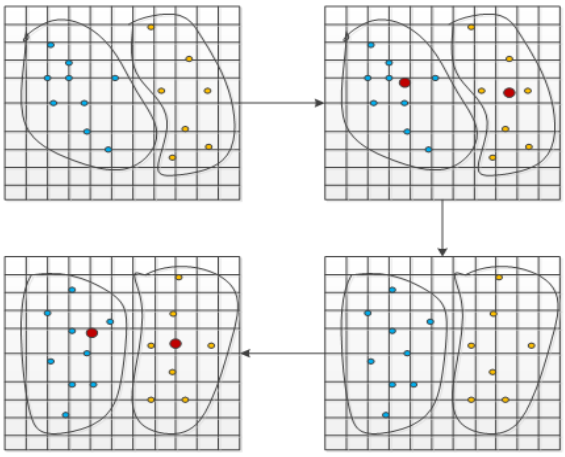

Figure 1. Process of cluster

$$\sum_{i=1, j \in \{1,2,K\}n} min \| x_i - p_j \|^2 \quad (1)$$

$k$ refers to the $k$ clustering centers, $x_i$ refers to the victor of clustering centers, $p_j$ refers to the victor of un-clustering centers. Renew the mean of the category:

$$\overline{x}_i = \frac{1}{|c_i|} \sum_{x \in c_i} x \quad (2)$$

$\overline{x_i}$ refers to the victor of clustering centers after renewing.

**3.2 Introduction of SAS**

SAS (Statistical Analysis System) was developed in 1966 by North Carolina State University as the statistical analysis software. SAS software institute (SAS Institute Inc.) was founded in 1967 to realize the deep processing research of the business data and the history data with the most advanced information and IT technology. SAS has an advantage in data mining and is widely used.

SAS has an enormous energy including the very effective apply in cluster analysis [6]-[8]. SAS can realize the function such as CLUSTER (hierarchical clustering), FASTCLUS (K-means clustering), MODECLUS (nonparametric clustering), VARCLUS (variables clustering), TREE. This paper takes the K-means.

**3.2.1 Process of the experiment**

**1. Steps**

The process of the experiment is figure 2:

a) Import the original viewing data and program data into SAS;

b) Retains the useful variables;

c) Relevant the viewing data and the program data by a corresponding relationship;

d) Calculate the duration to get the user duration victor set;

e) Divided users into different category

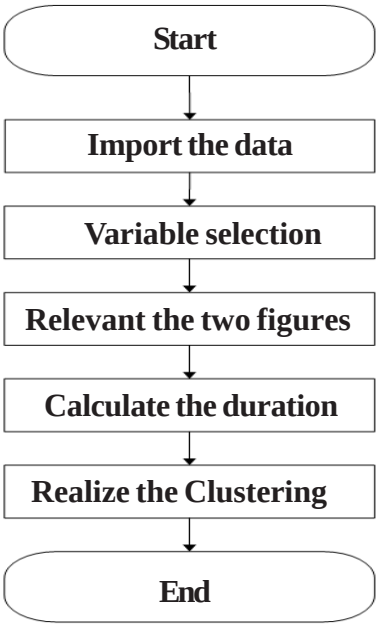


Figure 1. Processing of the experiment

**3.2.2 Data of the Experiment**

The data in this experiment is 17299 pieces user's digital TV viewing data and program data in 2013/02/01 from the provinces of China. The SAS version is 9.3.

There has a Software Development Kit(SDK) in the TV set-top box which can send back the viewing dat. The original data format is .txt which is as following:

**a) The Viewing Data:**

<GHApp>
<WIC
*cardNum* = "4271996831"
*stbNum* = "06111900501*C*60*B*8*EA*6800*A*933" *regionId* = "0*x*3010000"
*caArea* = "300"
*date* = "2013-02-01"
*pageWidgetVersion* = "1.0" >
<*A e* = "23:57:06"
*s* = "23:56:53"
*n* = "509"
*t* = "6"
*sn* = "*Jiangsu TV* " />
</WIC>
</GHApp>

*regionId* refers to the *region number*, *cardNum* refers to the TV set-top box number, *e* refers to the view start time, *s* refers to the view end time, *sn* refers to channel name.

**b) Program Data:**

171494|3748|21000109| Jiangsu TV ||||2013-2-01| 23:50:00|23:57:49|||AD|others|AD|||||||||2013-3-6 22:47:37|2013-3-6 22:56:05|kb|shenheall|pass||admin|| "21000109" refers to the channel number, "*Jiangsu TV*" refers to channel name, "2013-2-01" refers to the broadcast date, "23:50:00" refers to the broadcast start time, "23:57:49" refers to the broadcast end time, "*AD*" refers to the program type.

**3.2.3 data process**

SAS has its own data format [9]-[11], so, we should transform the original data into the SAS recognizable format. The data in SAS is as following:

**a) Viewing Data**

Pdbh refers to the channel number, cardNum refers to the TV set-top box number, StT refers to the view start time, SeT refers to the view end time, pdn refers to channel name.

| Pdbh | cardNum | StT | SeT | pdn | Date |
|---|---|---|---|---|---|
| 21000109 | 4271996831 | 23:56:53 | 23:57:06 | Jiangsu TV | 2013-02-01 |

Table 1. Viewing Data In SAS

**b) Program data**

| Pdbh | pdn | bcd | bct | jst | jmdl |
|---|---|---|---|---|---|
| 21000109 | Jiangsu TV | 2013-02-01 | 23:50:00 | 23:57:49 | Others |

Table 2. Program Data In SAS

Pdbh refers to the channel number, Pdn refers to the channel name, bcd refers to the broadcast date, bct refers to the program start time, jst refers to the program end time, Jmdl refers to the program type.

**3.2.4 Figures Relevant**

Using the joint variable to relevant the viewing figure and the program figure. After variable selection we can get the following table:

| cardNum | StT | SeT | pdn | bct | st | jmdl |
|---|---|---|---|---|---|---|
| 427199 6831 | 23: 56: 53 | 23: 57: 06 | Jiangsu TV | 23: 50: 00 | 23: 57: 49 | Others |

Table 3. Wide Table

**3.2.5 Duration Calculation**

The final victor is multidimensional, for example: zhang San( duration of program type 1, duration of program type 2, duration of program type 3…duration of program type *n*).

| Program \ User | User1 | User2 | … | User n |
|---|---|---|---|---|
| *C*1 | $T_{11}$ | $T_{12}$ | … | $T_{1n}$ |
| *C*2 | $T_{21}$ | $T_{22}$ | … | $T_{2n}$ |
| … | .. | .. | .. | .. |
| *Cm* | $T_{m1}$ | $T_{m2}$ | … | $T_{mn}$ |

The formula of duration ratio is as the following:

$$\text{Duration ratio} = \frac{\text{Viewing Duration}}{\text{Program Duration}} \tag{3}$$

There are 4 situations between the viewing time and the program time which is as the following:

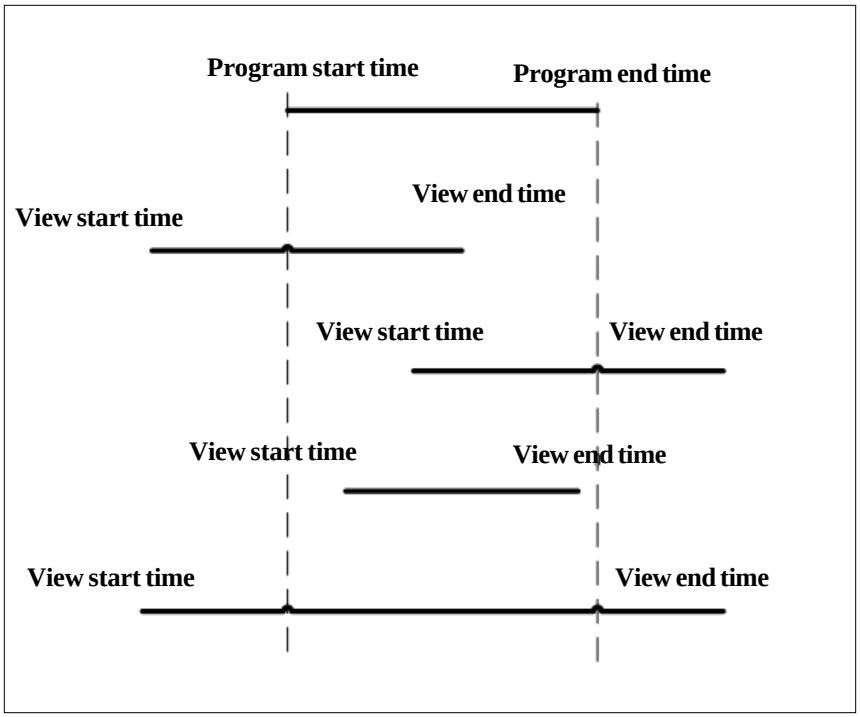


Figure 3.Relationships between viewing time and program time

Pro_StartTime refers to the program start time, Pro_EndTime refers to the program end time, Service_StartTime refers to the viewing start time, Service_EndTime refers to the program end time.

a) *if* Service_StartTime < = Pro_StartTime and Pro_StartTime < = Service_EndTime < = Pro_EndTime then *T*=Service_EndTime-Pro_StartTime;

b) *if* Pro_StartTime < = Service_StartTime < = ro_EndTime and Service_EndTime > = Pro_EndTime then *T*=Pro_EndTime-Service_StartTime;

c) *if* Service_StartTime > = Pro_StartTime and Service_EndTime < = Pro_EndTime then *T* = Service_EndTime-Service_StartTime;

d) *if* Service_StartTime < = Pro_StartTime and Service_EndTime > = Pro_EndTime then *T* = Pro_EndTime-Pro_StartTime;
Based on the above formula we can get the following figure:

From the above figure we can see that the *cardnum* refers to different users and we put the TV shows into 8 different types which are life service type, TV play type special subject type, live show type, movie type, education type, news type, others type. For example: user 4153019302, its viewing track is that the duration of the life service type programs is 48 units, the duration of the TV play type programs is 10.85 unit, the duration of the special subject type programs is 46.48 units, the duration of the live show type programs is 47 units, the duration of the movie type programs is 0 units, the duration of the education type programs is 0.18 units, the duration of the news type programs is 0 units, the duration of the other type programs is 48 units.

| cardnum | life service | TV play | special subject | live show | movie | education | news | others |
|---|---|---|---|---|---|---|---|---|
| 4153019302 | 48 | 10.85 | 46.48 | 47 | 0 | 0.18 | 0 | 48 |
| 4153019320 | 0 | 0 | 0 | 0 | 0.83 | 0 | 0 | 0 |
| 4153022954 | 45 | 42.89 | 45 | 45 | 0 | 0 | 45 | 0 |
| 4153023170 | 0 | 0 | 0 | 0.64 | 0 | 0 | 0 | 0 |
| 4153023749 | 72 | 72 | 72 | 71.76 | 0 | 0 | 72 | 0 |

Figure 4. Duration ratio

**3.3 Result of Experiment**

Firstly, we have the data normalized and the result is as the following:

| cardnum | life service | TV play | special subject | live show | movie | education | news | others |
|---|---|---|---|---|---|---|---|---|
| 4153019302 | 1.1 | -0.02 | 0.997 | 1.936 | 0 | -0.52 | 0 | 0.71 |
| 4153019320 | 0 | 0 | 0 | 0 | -0.37 | 0 | 0 | 0 |
| 4153022954 | 0.9782 | 1.542 | 0.939 | 1.833 | 0 | 0 | 1.2 | 0 |
| 4153023170 | 0 | 0 | 0 | -0.45 | 0 | 0 | 0 | 0 |
| 4153023749 | 2.0831 | 2.956 | 1.933 | 3.206 | 0 | 0 | 2.4 | 0 |

Figure 5. Data normalized

Setting *k* = 4, and the maxiter = 100 we can get the following result:

| cluster | life service | TV play | special subject | live show | movie | education | news | others |
|---|---|---|---|---|---|---|---|---|
| 1 | 8.16 | 4.57 | 7.84 | 10.91 | 0 | -0.19 | 8.8 | 0 |
| 2 | 6.59 | 8.3 | 6.25 | 8.71 | 0 | 9.21 | 7.1 | 5.58 |
| 3 | 3.54 | 4.58 | 3.39 | -0.41 | 0 | 5.27 | 3.9 | 3.92 |
| 4 | -0.86 | -0.54 | -0.81 | 0 | 0 | 0 | -1 | -0.63 |

Figure 6. Original cluster center

In the processing of cluster, the most important value is the mean value of cluster [12]. We can get user's interests from the biggest mean value of a category which indicated the user's interest is the corresponding program. The mean value of cluster is as the following:

| cluster | life service | TV play | special subject | live show | movie | education | news | others |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.9871 | 0.9852 | 0.967 | 0.5496 | 0.9912 | 0.0677 | 0.9727 | 0.9282 |
| 2 | 0.1199 | 0.0785 | 0.1152 | 0.0648 | 0.1114 | 0.1427 | 0.1175 | 0.178 |
| 3 | 0.8881 | 0.1844 | 0.8102 | 0.9238 | 0.8572 | 0.3461 | 0.8207 | 0.821 |
| 4 | 0.9703 | 0.9916 | 0.9513 | 0.9751 | 0.0695 | 0.9732 | 0.9774 | 0.9846 |

Figure 7. Mean value of cluster

Taking the fourth class as an example. The biggest value is 0.9913, so, users who belong to the fourth class like TV play most.

SAS gives us the cluster result as a table that can be exported as .txt. The following figure is the result:

| cardnum | life service | TV play | special subject | live show | movie | education | news | others | cluster |
|---|---|---|---|---|---|---|---|---|---|
| 4153019302 | 1.1 | -0.02 | 0.997 | 1.936 | 0 | -0.52 | 0 | 0.71 | 3 |
| 4153019320 | 0 | 0 | 0 | 0 | -0.37 | 0 | 0 | 0 | 4 |
| 4153022954 | 0.9782 | 1.542 | 0.939 | 1.833 | 0 | 0 | 1.2 | 0 | 3 |
| 4153023170 | 0 | 0 | 0 | -0.45 | 0 | 0 | 0 | 0 | 4 |
| 4153023749 | 2.0831 | 2.956 | 1.933 | 3.206 | 0 | 0 | 2.4 | 0 | 1 |

Figure 8. Cluster Result

To get the best result, we set the contrast experiment by change the value of the value of *k* and the maxiter. By contrasting the different experiment set, we find that $k = 4$ is the best setting.

This experiment achieves the desired affects.

**Acknowledgment**

This paper is supported by the national sci-tech support plan project Key Technology Research and Application of Stereo Vision Content Services which numbered 2012BAH37F03. Thanks for the work of the refereeing.

**References**

[1] Liuxi, Wei., Ronghuan, Lin. (2012). A Method for Customers Division of Device Management Platform Using SPRINT Decision. Software,11

[2] Slanjankic, E., Balta, H., Joldic, A. (2009). Data mining techniques and SAS as a tool for graphical presentation of principal components analysis and disjoint cluster analysis results. information, communication and automation technologies.

[3] Maj., Liao J x., Zhu X m. (2008). Device management in the IMS.J. Journal of Network and Systems Management, 16(1): 46-62.

[4] Xiang liang. Recommendation system practice. Beijing, posts and telecom press, 2012.6

[5] (Sundar) Balakrishnan, P. V. (1995). Nicholas G. Hall, A maximin procedure for the optimal insertion timing of ad executions, European Journal of Operational Research 85, 368–382.

[6] Peter, J., Danaher., Roland, T., Rust. (1996). Determining the optimal return on investment for anadvertising campaign, European Journal of Operational Research 95, 511–521.

[7] Harvey, B. (1968). Nonresponse in TV meter panels. Journal of Advertising Research, 8:24–7.

[8] Mayer, M. (1965). How good are television ratings?. New York: Television Information Office.

[9] Buck, S. (1983). TV audience research-present and future. Admap, December, p.636–40.

[10] Gensch, D., Shaman, P. (1980). Predicting TV ratings. Journal of Advertising Research, August:85–92.

[11] Horen, J. (1980). Scheduling of network television programs. Management Science, April:354–70

[12] Lovejoy, W. (1987). Ordered solutions for dynamic programs. Mathematics and Operations Research, 12:269–76.

[13] Neuman, W. (1997). Social research methods: qualitative and quantitative approaches. Needham Heights, MA: Allyn & Bacon