Predicting Television Ratings and Its Application to Taiwan Cable TV Channels

Hui-Ling Huang
Department of Biological Science and Technology,
National Chiao Tung University
Institute of Bioinformatics and Systems Biology,
National Chiao Tung University
Hsinchu, Taiwan
hlhuang@mail.nctu.edu.tw

Hua-Chin Lee
Department of Biological Science and Technology,
National Chiao Tung University
Institute of Bioinformatics and Systems Biology,
National Chiao Tung University
Hsinchu, Taiwan
huachinlee@nctu.edu.tw

Li-Sun Shu
Department of Multimedia and Game Design,
Overseas Chinese University
Taichung, Taiwan
Issu@ocu.edu.tw

Shih-Chung Lai
Institute of Bioinformatics and Systems Biology,
National Chiao Tung University
Hsinchu, Taiwan
ratadune@gmail.com

Tse-Ming Tsai
Institute for Information Industry
Taipei, Taiwan
eric@iii.org.tw

Shih-Chun Chou
Institute for Information Industry
Taipei, Taiwan
benchou@iii.org.tw

Bo-fu Liu
Institute for Information Industry
Taipei, Taiwan
bofuliu@iii.org.tw

Yun-Ju Yin
Institute of Bioinformatics and Systems Biology,
National Chiao Tung University
Hsinchu, Taiwan
yunjuyin@gmail.com

Hong-An Chen
Institute of Bioinformatics and Systems Biology,
National Chiao Tung University
Hsinchu, Taiwan
ax428285@gmail.com

Shinn-Ying Ho*
Department of Biological Science and Technology,
National Chiao Tung University
Institute of Bioinformatics and Systems Biology,
National Chiao Tung University
Hsinchu, Taiwan
syho@mail.nctu.edu.tw
*Corresponding author

***Abstract*—Using forecast television network ratings, television executives estimate a price to sell time to advertisers. TV rating is an important feedback mechanism because its results greatly affect the immense profits of TV companies, advertisers, and program producers. Therefore, how to select the samples for TV rating investigation plays an important role in predicting program ratings. How to design an accurate predicting model for program rating also is an important investigation. The predicting problem is essentially a bi-objective optimization problem which minimizes the number of samples and maximizes the predicting accuracy of program rating. In this study, we propose an evolutionary approach to designing a rating model (ERM) by simultaneous optimization of sampling sub-area selection and parameter tuning using an intelligent genetic algorithm (IGA). In this study, the ERM is applied to Taiwan Cable TV Channels in Taipei and Taiwan. The experiments show that TV rating prediction of the proposed ERM is efficient smaller than that of using the same number of sub-areas with the largest TV ratings and an optimal prediction program rating by using the selected sub-areas.**

Keywords-component; formatting; TV rating; digital set-top-box; sampling metho; IGA; rating model

I. INTRODUCTION

TV rating has long been seen as an important feedback mechanism. Its results greatly affect the immense profits of

TV companies, advertisers, and program producers, and also determine the length of the programs. In terms of suppliers, since the development of hardware in the audience rating industry, such as telecommunication apparatuses and individual recording machines, have become stabilized, the suppliers of such hardware are lacking in bargaining power. Conversely, once the digital set-top-box becomes the viewing method for the audience rating market, its bargaining power will increase.

Program rating with advertising impact has become a key factor in the value of time. Developmental audience rating prediction is in dire need. Review past research in audience appreciation and develop a metric from the ratings data that can be used as an early predictor of program growth or decline [1]. Linear regression models for the prediction of population rating have used by several authors [2]. Therefore, collecting the dataset for producing accurate rating predictions is a challenging task. One of the most challenging tasks of collecting data is finding target audience from big data flow. The majority of performance ratings of the choosing sample size can still closely represent entirety program ratings.

The existed studies for sampling methods are described as below. Breiman et al. [3] used F-tests to determine the optimum splitting points for decision tree models. Hartigan [4] first published CHAID decision trees to obtain the optimum splitting points with chi-squared.

Ho et al. [5] proposed intelligent evolutionary algorithms based on orthogonal experimental designs for solving large parameter optimization problems. The intelligent genetic algorithm (IGA) is one customized version of the intelligent evolutionary algorithm for solving specific problems.

We propose an evolutionary approach to designing a rating model (ERM) by simultaneous optimization of sampling sub-area selection and parameter tuning using IGA. In this study, the provided audience behavior data including 140 million data records and 185 thousand peoples was collected from Taiwan digital CATV system in January of 2013. The Taiwan region has 125 sub-areas and the Taipei region (the capital in Taiwan) has 29 sub-areas. In this study we apply Stylish Man-The Chef (11:00~13:00) program to the proposed ERM for prediction from 2013/1/14~18 totally n=5 day's ratio of channel viewing times.

II. METHOD

The ability of ERM arises mainly from simultaneous optimization of parameter setting of rating model and sampling rating sub-areas selection using IGA. IGA adopts an efficient GA-chromosome encoding scheme. ERM may select a minimum number of sub-areas from statistics of the TV dataset and maximum closed programing ratings. It is intractable to simultaneously optimize the two objectives.

A. IGA for ERM

ERM makes the best use of IGA in optimizing system parameters by designing an efficient GA-chromosome representation as well as an intelligent crossover operation. The intelligent crossover operation is based on an orthogonal experimental design using a divide-and-conquer strategy to solve intractable optimization problems comprising lots of system parameters.

- *Fitness function and GA-chromosome representation*

Fitness function is the only guide for IGA to optimize all system parameters encoded into a GA-chromosome. There are three objectives for designing ERM using IGA. The first is to minimize an error value which responds to sampling data how far away from the real channel viewing time's C_a, the second is to minimize the number N_f of selected sub-areas and the third is minimize total error of prediction program rating of the next day P_r.

$$C_a=\sum_{d=1}^{n}\left|R_k^d - S_k^d\right| \tag{1}$$

where R_k^d is the d^{th} day of real channel viewing times, S_k^d is the d^{th} day of sampling channel viewing times, n is number of days, N is number of sub-areas, k=1,.., N.

$$P_r=\sum_{d=2}^{n}\left|R_{next}^d - M_{next}^d\right| \tag{2}$$

where R_{next}^d is the $(d+1)^{\text{th}}$ day of real different ratio of channel viewing times, M_{next}^d is the $(d+1)^{\text{th}}$ day of ERM model ratio of prediction of channel viewing times. The equations of R_{next}^d and M_{next}^d are shown as follows:

$$R_{next}^d = \frac{(S_k^d - S_k^{d-1})}{S_k^d} \tag{3}$$

$$M_{next}^d = \alpha\left(\frac{(S_k^{d-1} - S_k^{d-2})}{S_k^{d-1}}\times 100\ \%\right) + \beta \tag{4}$$

where α, β are the parameters of ERM which are encoding in GA-chromosome. If S represents the set of parameters to be evolved by IGA, the fitness function $y(S)$ is as follows:

$$\text{Min } y(S) = C_a(S)+w_1N_f(S)+w_2P_r(S) \tag{5}$$

where w_1 is a positive weight determined according to the preference of individual objectives. Generally, high program rating of the next day is the major objective and thus w_2 is set to a small constant value.

Let $S = \{t_i, \alpha, \beta | i=1,2,\ldots,l\}$ encoded into a GA-chromosome. The control GA-genes $t_i \in \{0,1\}$ are used to select effective sub-areas. α and β are the control parameters of ERM, and l is a prespecified maximum number of selected sub-areas.

- *Intelligent genetic algorithm(IGA)*

The used IGA for ERM to optimize the parameters in S using the fitness function $y(S)$ is given as follows:

- Step 1: Initialization. Randomly generate an initial population with N_{pop} feasible individuals where each gene g_i is unique in a GA-chromosome.
- Step 2: Evaluation. Evaluate fitness values of all individuals in the population. Let I_{best} be the best individual in the population.
- Step 3: Use the simple truncation selection that replaces the worst $P_s \bullet N_{pop}$ individuals with the best $P_s \bullet N_{pop}$ individuals to form a new population, where P_s is a selection probability.
- Step 4: Randomly select $P_c \bullet N_{pop}$ individuals

including I_{best}, where P_c is a crossover probability. Perform intelligent crossover operations for all selected pairs of parents.

- Step 5: Apply a conventional bit-inverse mutation operator to the population using a mutation probability P_m. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.
- Step 6: Termination test. If a prespecified termination condition is satisfied, stop the algorithm. Otherwise, go to step 2.

B. DATASET

The dataset was collected from Taiwan digital CATV system in January of 2013 and was supported by the Institute for Information Industry (III). In order to further process the suitable and usable dataset, the wrong data format and inconsistent data was filtered out. In addition, the unsuitable TV program receiving time was also eliminated by checking if the time is less than 5 minutes or more than 6 hours. Finally, the raw dataset remains 14 dimensions of information such as the smartcard id of Set-top box (STB), user location, user behavior type and behavior happen time & end time. Besides, the Electronic Program Guide (EPG) of January 2013 additionally provides the information of program id, program name, program start/end time and program type.

TABLE I THE DESCRIPTIVE STATISTICS OF THE DATASET

Region	data record (unit: million)	Sub-area	People (thousand)
Taipei region: Taipei and new Taipei city	35	29	76
THM region: Taoyuan-Hsinchu-Miaoli region	10	8	23
TCN region: Taichung-Changhua-Nantou region,	8	22	19
YCT region: Yunlin-Chiayi-Tainan region,	3	17	5
KP region: Kaohsiung-Pingtung region,	9	39	20
YHT region: Yilan-Hualien-Taitung region,	2	10	4

Table 1 shows the descriptive statistics of the dataset in different areas. There are 221 channels which include 37 entertainment channels in Taiwan digital TV system, and the 125 sub-areas contain 140 million data records from 185 thousand people.

C. Preprocessing

Step1: Filtering. This dataset contains 140 million data records and 185 thousand users. The first step is filtering the invalid data record. (1) Filter that data record which the *channel watching time* is less than 30 seconds. (2) Filter that user which watching day is less than 3 days in one month's time. After Filtering, the new dataset contains 67 million data records and 147 thousand users.

Step2: Data divided. This step is divided the huge dataset (totally 125 sub-areas) into 6 regions, which responses to the administrative area. The 6 areas are: (1) Taipei region (2) THM region (3) TCN region (4) YCT region (5) KP region (6) YHT region

Step3: Data statistical analysis. This study analyzes the data records from January 14th to January 18th. (1) Choosing the comprehensive programs from totally 221 channels. (2) Summarize the viewing times for each comprehensive program channel.

Channel selection. In Taiwan, most of TV Channels prefer the analysis of the business-orientated information about the application of TV Rating. Viewing time has resulted in affecting gigantic profits for TV companies, advertisers, and program producers, with the data offered. This study chooses Sanlih Entertainment Television City which is the highest viewing times channel to analyze and the Statistical chart is seen in Fig 1. By analyzing viewing times, it will represent the channel viewership.

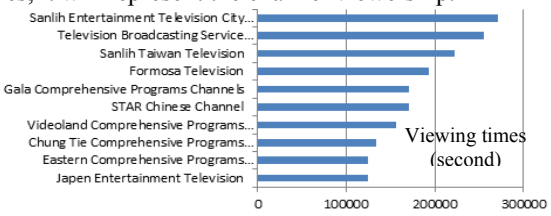

Fig. 1 Viewing Times Top10 Channels in 2013/1/14~18

Time zones selection. Due to the high ratings will be able to create high-priced advertising revenue, TV companies will arrange high-attractive program at high viewing times. However the kind of program also affect the different high viewing times. Program and time interacts viewership with each other. This study chooses *Stylish Man-The Chef (11:00~13:00) program* which is the highest viewing program to analyze and the Statistical chart is seen in Fig.2.

D. Method using ERM

In this study we apply *Stylish Man-The Chef (11:00~13:00) program* to the proposed ERM for prediction from 2013/1/14~18 totally n=5 day's channel viewing times.

1) Majority region choosing. This study selects Taipei region which include Taipei the Capital city (l = 29) and whole Taiwan region (l = 125) these two different ranges to be the sampling models.

2) ERM: The parameter settings of IGA are N_{pop} = 30, P_c = 0.5, P_s = 0.2 and P_m = 0.2. Let the weighting $w_1 = 1/l$ in the fitness function $y(S)$. The stopping condition of IGA is to use 100 generations. Because of the non-deterministic characteristic of GA, R = 30 independent runs are performed.

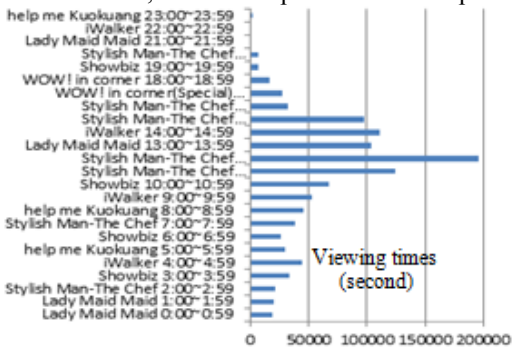

Fig.2 Sanlih Entertainment Television Viewing Times in 2013/1/14-1/18

III. RESULTS

A. *Training results by Taipei region ERM*

The program ranting data from Taipei areas, there are 29 sub-areas data to be created ERM for predicting the future viewing times ratio. The encoding length l =29, the objective function w_1 = 1/29, w_2 = 10, the IGA parameters are setting. The number of selected sub-areas N_f =10, the selected sub-areas are ShihlinDist, Datong Dist, Beitou Dist, Songshan Dist, Sanjhih Dist, Sanchong Dist, Wugu Dist, Shenkeng Dist, Sindian Dist and Sinjhuang Dist. The ERM parameters α and β are 0.0158 and -2.2653, respectively.

TABLE II . 1/14-18 REAL AND PREDICTING RATIO OF VIEWING TIMES.

	2013/1/14	2013/1/15	2013/1/16	2013/1/17	2013/1/18
real area (viewing times)	3029601	2421220	2381400	3800251	3475044
growing rate(%)	0	-20.0812	-1.6446	59.5805	-8.5575
predict growing rate(%)			-2.5828	-2.2913	-1.3234
predict error rate(%)			-0.9382	-61.8718	7.2341

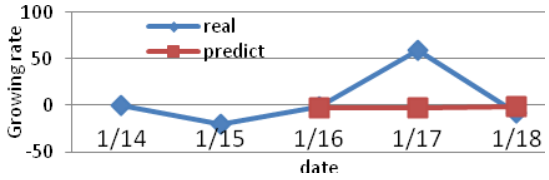

Figure 3. The growing rate between real and predict viewing times value.

Appling the ERM to predicting ratio viewing time of next day, the objective function P_r needs two days before real ratio of viewing time. For example, according the date of 2013/1/14~15 to predict the ratio of viewing times of the date of 2013/1/16. Table 2 and Fig. 3 show the results of training ERM. The blue line is real viewing times ratio, and red line is the predicting next ratio of viewing times in Fig. 3.

B. *Independeent test results by Taipei region ERM*

Using the next week data from 2013/1/21~25 to be independent test ERM, the predicting ratio of viewing time is shown in Table 3 and Figure 4. The blue line is real ration of viewing times, and red line is the predicting ration of viewing times. The blue line is use to be a control group and compare with the red line to reflect error values.

TABLE III. 1/21-25 REAL AND PREDICTING RATIO OF VIEWING TIMES.

	2013/1/21	2013/1/22	2013/1/23	2013/1/24	2013/1/25
real area (viewing times)	3357279	3194936	2933051	2889403	3082288
grow rate(%)	-3.3889	-4.8356	-8.1969	-1.4881	6.6756
predict growing rate(%)	-2.4006	-2.3189	-2.3418	-2.3949	-2.2888
predict error rate(%)	0.9883	2.5167	5.8551	-0.9068	-8.9644

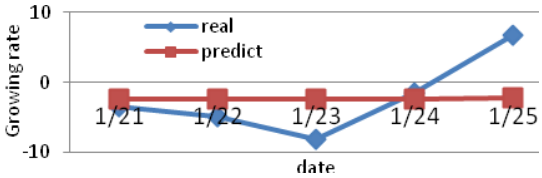

Figure 4. The growing rate between real and predict viewing times value.

C. *Training results by the whole Taiwan ERM*

The IGA parameters are setting as follows: the encoding length l =125, the objective function w_1 = 1/125, w_2 = 10. The predicting results of ERM in the whole Taiwan region are as follows: the number of selected sub-areas N_f is 45, the ERM parameters α and β are 0.0002 and -0.0155, respectively.

Appling the whole Taiwan region ERM to predicting ratio viewing time of next day, the objective function P_r needs two days before real ratio of viewing time. Table 4 and Fig. 5 show the results of training ERM. The blue line is real viewing times ratio, and red line is the predicting next ratio of viewing times in Fig. 5.

TABLE IV. 1/14-18 REAL AND PREDICTING RATIO OF VIEWING TIMES.

	2013/1/14	2013/1/15	2013/1/16	2013/1/17	2013/1/18
real area (viewing times)	5706803	4655241	4654006	7181867	6835830
growing rate(%)	0	-18.4264	-0.0265	54.3158	-4.8182
predict growing rate(%)			-0.0193	-0.0155	-0.0043
predict error rate(%)			0.0072	-54.3313	4.8139

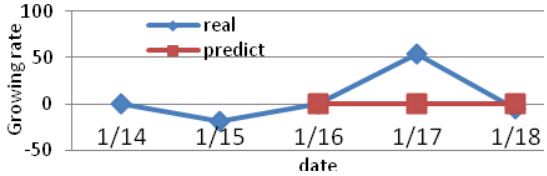

Figure 5. The growing rate between real and predict viewing times value.

D. *Independeent test results by the whole Taiwan ERM*

Using the next week data from 2013/1/21~25 to be independent test ERM, the predicting ratio of viewing time is shown in Table 5 and Figure 6. The blue line is real ration of viewing times, and red line is the predicting ration of viewing times. The blue line is use to be a control group and compare with the red line to reflect error values.

TABLE V. 1/21-25 REAL AND PREDICTING RATIO OF VIEWING TIMES.

	2013/1/21	2013/1/22	2013/1/23	2013/1/24	2013/1/25
real area (viewing times)	7291149	6715333	7045359	6328426	6597261
grow rate(%)	6.6607	-7.8975	4.9145	-10.1760	4.2480
predict growing rate(%)	-0.0165	-0.0141	-0.0171	-0.0145	-0.0176
predict error rate(%)	-6.6772	7.8834	-4.9316	10.1615	-4.2656

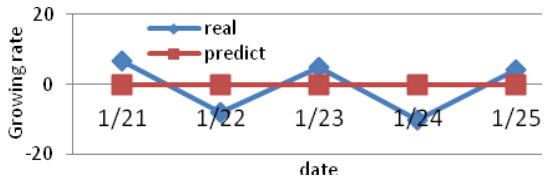

Figure 6. The growing rate between real and predict viewing times value.

IV. CONCULSION

This paper has proposed an efficient evolutionary predicting model named ERM by the intelligent genetic algorithm (IGA). The automatic sub-area selection and parameter tuning embedded in ERM are simultaneously optimal by IGA, which can advance the predicting model performance from a large number of digital set-top-box data.

ERM can serve not only as a program rating predictor but also as an adaptive sampling extractor. ERM is developed as an efficient tool of program rating predictor.

V. ACKNOWLEDGMENT

This study is conducted under the "Digital Convergence Key Technology and System Develop Project" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China. The authors would like to thanks the member of the Innovative DigiTech-Enabled Applications & Services Institute at the Institute for Information Industry (III) New Media User Lab, and their collaborators, whose contributions were central to this paper.

REFERENCE

[1] V. Beal, A. Tanusondjaja, M. Nenycz-Thiel. Behavioural Measures: Can they predict television program growth and decline?. The Australian and New Zealand Marketing Academy Conference 2011 Perth, Australia 2011-11-28.

[2] D. Meyer and R. J. Hyndman. The Accuracy of Television Network Rating Forecasts: The Effects of Data Aggregation and Alternative Models. MASA, 1(3), 145-154, 2006.

[3] L. Breiman, J. Friedman, R. Olshen and C. Stone. Classification and Regression Trees, Wadsworth, 1984.

[4] J.A. Hartigan, Clustering Algorithms,Wiley: NewYork, 1975.

[5] S.-Y. Ho *et al.*. Intelligent evolutionary algorithms for large parameter optimization problems. *IEEE Trans. Evol. Comput.*, **8**, 522-541,(2004b).