ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 682-686
doi: 10.18178/wcse.2016.06.121

Selecting Classification Model for the Personalized Movie Recommendation System by Feature Adjustment Method

Supachanun Wanapu [1+], Thawatphong Phithak [2] and Narodom Kittidachanupap [3]

[1] Faculty of Management Science, Nakhon Ratchasima Rajabhat University, Nakhon Ratchasima, Thailand
[2] School of Information Technology, Suranaree University of Technology, Nakhon Ratchasima, Thailand
[3] Faculty of Science Technology and Agriculture, Yala Rajabhat University, Yala, Thailand

Abstract. Recommendation systems are widely used to improve market potential in theater business today. However, the efficiency of the personalized movie recommendation system (PMRS) using model-based techniques is related to employed classifier and a number of features. This research aims to select a suitable classification model by feature adjustment method for creating the recommendation rules of PMRS. The suggestion model is appraised using retrieval performance measures by Accuracy between 3 algorithms of classification consisting of J48, Naïve Bayes (NB) and Multilayer Perceptron (MLP). The datasets for model construction are collected through surveying from 383 movie audiences who live in Nakhon-Ratchasima province, Thailand. The results of the accuracy performance show that J48 algorithm produces the finest accuracy (70.28%) followed by NB (68.28%) and MLP (66.23%), respectively. In addition, the performance of J48 by feature adjustment method provides 58 combinations which are created from 6 features of movie audience's profile and 19 features of movie genres. The results of feature adjustment method present the consistency between accuracy performance and a number of features. However, the progress of recommendation rules set selection for PMRS will be chosen only 36 high performance combinations of adjustment features and these combinations will be applied to the development of a new personalized movie recommendation system.

Keywords: personalized recommendation, classification, feature adjustment.

1. Introduction

Theater business in Thailand has expanded and grown continually. Many entrepreneurs have modernized strategic plan for delivering several services to customers such as building the theater for 3D and 4DX movies, developing modern products and more comfortable additional services, etc. For this reason, the recommendation systems for theater business are developed. It is used to release the movie news and promotions, which are suitable for particular customers. Moreover, the recommendation systems enlarge the market potential in the theater business.

In the past, the development of a movie recommendation systems by data mining procedure were based on Knowledge-Based Systems (KBSs) techniques. The received information from KBSs will be used to perform the suggestions. Currently, the Model-based techniques are interested for building movie recommendation systems because this technique will create a proper model before recommendation process. The output model will be used to evaluate the suggestion swiftly. Moreover, the use of appropriate classification algorithms can be provided the target results explicitly [1-3]. However, various algorithms for model construction in the data mining have some problems. Examples of such algorithms are Artificial Neural Networks (ANN), Support Vector Machines (SVMs), Naïve Bayes and Decision Tree (DT). ANN is a popular approach widely used to solve classification problems. However, ANN's relative importance of potential input variables, long training processes, and interpretative difficulties have often been criticized.

[+] Corresponding author. Tel.: +66 81 548 6455.
E-mail address: supachanun.w@nrru.ac.th.

SVM has high performance in classification problems. However, the rules obtained by the SVM algorithm are hard to understand directly [4]. DT is a basic form of supervised learning and it represents one of the most popular approaches for classification problems. However, a disadvantage of DT is that it only handles discrete attributes and it does not allow multiple output attributes [5-6]. If the relationships between input features are weak, DT may provide poor classification accuracy [7, 8]. These problems occur because unsuitability of classifier and the number of features on datasets matrix.

This work is preliminary progress of model-based recommendation. The objective of this research is to adjust the proper number of features for selecting classification models. The optimal results will be created the recommendation rules set, which is some part of the development of the Personalized Movie Recommendation System (PMRS). The explanation of this research is divided into 5 sections. Section 2 presents the proposed PMRS architecture. Section 3 describes the matrix of datasets in the model construction. Section 4 informs the experimental environment and results. The last section, Section 5 reports the conclusion and future work.

2. PMRS Architecture

The proposed Personalized Movie Recommendation System (PMRS) is established in order to find the correlation between the related factors of movie audience’s characteristics and movie genres. The PMRS model is designed and developed using a modular approach, which is divided into 4 stages consisting of 1) Model-based recommendation modules, 2) Knowledge base, 3) Personalized search module and 4) System GUI. Fig. 1 shows the architecture of the PMRS system, which expose an overview of the PMRS system components and its data flow.

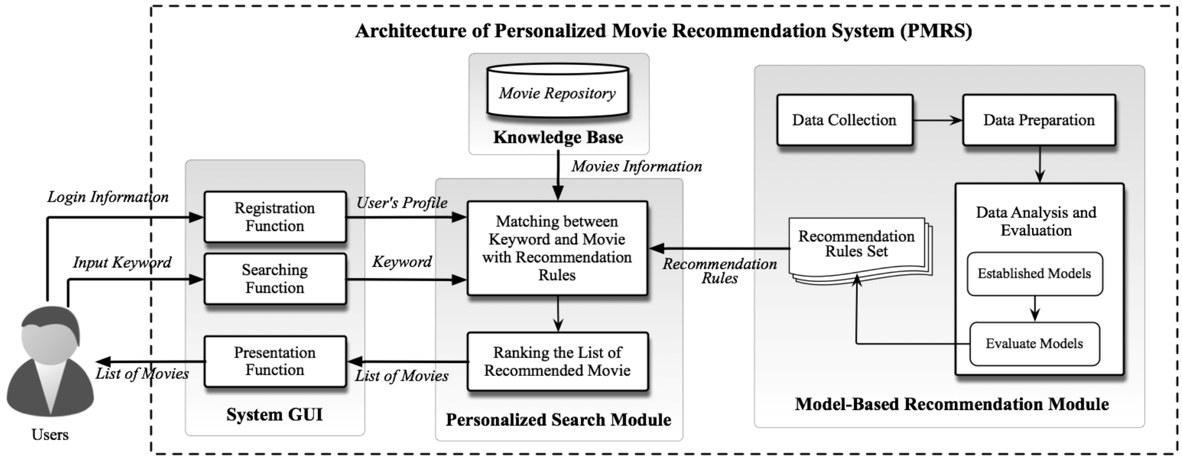

Fig. 1: The architecture of personalized movie recommendation system (PMRS).

The process of PMRS starts when users log in and send the registration information to the system via system GUI. Thereafter, when users input keyword to the system, a personalized search module will be initiated to collect the user's profile and keyword for the matching process. The matching process performs by finding the correlations between user's profile and recommendation rules which are received from the model-based recommendation module. The results of matching process will be used to search the movie information in the knowledge base. Then, PMRS will rank the list of recommended movies which are suitable for each user. However, the aim of this study is to develop a methodology to find the recommendation rules, which is preliminary progress of PMRS. The model-based recommendation module initiates from data collection and data preparation stage. Then, the completed data will be operated in data analysis and evaluation stage for establishing proper models and building recommendation rules set of PMRS.

3. Matrix of Datasets

The datasets for model construction are collected through surveying from the general public who live in Nakhon-Ratchasima province which has the second largest population of Thailand after Bangkok. The sample size is 383 movie audiences which are estimated by using Probability Proportional to Size (PPS) calculation. The surveying performs by using short-term questionnaire which is separated into two question sections as follows:

1) Profiles of movie audience consist of 10 features: *Gender* {male, female}, *Age* {lower 15, 16-20, 21-25, 26-30, 31-35, over 36}, *Occupation* {student, employee, general officer, business owner, home maid, other}, *Day* {weekday, weekend, holiday, uncertain}, *Frequency* {infrequent, 1-2 times/month, 3-4 times/month, 5-6 times/month, over 6 times/month}, *Education* {primary school, high school, bachelor, graduate, other}, *Status* {single, married}, *Salary* {lower 5000 THB, 5001-10000 THB, 10001-20000 THB, over 20000 THB}, *Ticket Type* {via online tickets website, via automatic ticket machine, via ticket counter} and *Payment Type* {counter service, credit card, m-cash, m-pay}.

2) Favourite movie genres consist of 19 features: Action, Thai-film, Fantasy, Animation, Crime, Film-Noir, Romance, Erotic, Sci-Fi, Musical, Comedy, Western, Documentary, War, Drama, Mystery, Thriller, Adventure and Family. With this question, participants can choose more than one answer.

After the data collection, the data preparation is performed. The data preparation employs an Apriori algorithm for analyzing the correlation of each feature. The Apriori algorithm is determined based on a minimum support threshold of 0.9, the results show only 6 features which are related to movie genres consisting of Age, Education, Status, Salary, Ticket Type and Payment Type. These features will be used to classify the model in the next section. Table 1 illustrates an example matrix of the datasets.

Table 1. An example of datasets for model construction

Profile of movie audience							**Favourite movie genre**				
ID	**Age** (A)	**Education** (B)	**Status** (C)	**Salary (THB)** (D)	**Ticket Type** (E)	**Payment Type** (F)	**Action**	**Thai-film**	**Fantasy**	**...**	**Family**
1	16-20	High School	Single	5001-10000	Counter	Counter Ser.	N	N	N	...	Y
2	Over36	Bachelor	Married	Over 20001	Counter	Credit Card	Y	N	Y	...	N
3	Over36	High School	Married	10001-20000	Machine	Counter Ser.	N	Y	N	...	N
4	Lower15	High School	Single	Lower 5000	Online	M-Cash	N	N	Y	...	N
5	21-25	Bachelor	Single	5001-10000	Counter	Credit Card	Y	N	Y	...	N
...	...	...	...	...	...	...	...	...	...	...	...
383	21-25	Bachelor	Single	10001-20000	Counter	M-Cash	N	N	N	...	N

4. Experimental and Result

The experiment employs 383 records on dataset; each record consists of 25 features (6 features from significant movie audience's profile and 19 features from movie genre).The preparatory evaluation of model uses the WEKA (The Waikato Environment for Knowledge Analysis) and 10-fold cross validation [9] on the training dataset. The model is appraised using retrieval performance measures by Accuracy between 3 algorithms of classification consisting of J48, Naïve Bayes (NB) and Multilayer Perceptron (MLP). The objectives of this experiment consisting of 1) Selecting suitable algorithm for the dataset, 2) Comparing the efficiency of the best classification algorithm by feature adjustment method, and developing the recommendation rules set for PMRS.

4.1. Suitable Classification Algorithm

The results of the accuracy performance show that the maximum accuracy found in the *Status* classifier. Moreover, J48 produces the best accuracy (70.28%) followed by NB (68.28%) and MLP (66.23%), respectively as shown in Table 2.

Table 2. The evaluation results of recommendation rules set

Classifier **Accuracy (%)**	Age (A)	Education (B)	Status (C)	Salary (D)	Ticket Type (E)	Payment Type (F)	**Average**
J48	70.76	67.62	**84.33**	53.00	71.80	74.15	**70.28**
NB	73.11	61.62	**77.28**	54.83	71.80	71.02	68.28
MLP	69.97	59.27	**80.68**	51.44	67.36	68.67	66.23

4.2. Performance of Decision Tree by Feature Adjustment Method

As explained in the previous section, J48 creates the best accuracy of the recommendation model. This section uses 6 features of user's profile as classifier. Then, the feature adjustment method is operated for summarizing the finest feature set. The feature adjustment method performs by creating the combination between 6 features of user's profile (based features) and 19 features of movie genre. It makes 58 combinations of classifier which are divided into 6 categories consisting of 20 features set (1+19), 21 features (2+19), 22 features set (3+19), 23 features set (4+19), 24 features set (5+19) and 25 features set (6+19). This adjustment method makes the significant performance as shown in Table 3.

Table 3. The evaluation results of feature adjustment method

Combination No.	Number of Features	Input Features Adjustment	Classifier of Classification Models												Average		
			Age (A)		Education (B)		Status (C)		Salary (D)		Ticket (E)		Payment (F)				
			Accuracy (%)	Improvement (%)	Accuracy (%)	Improvement (%)	Accuracy (%)	Improvement (%)	Accuracy (%)	Improvement (%)	Accuracy (%)	Improvement (%)	Accuracy (%)	Improvement (%)	Accuracy (%)	Improvement (%)	
1	**20**	**A,B,C,D,E,F**	**46.74**	-	**57.44**	-	**73.89**	-	**40.99**	-	**68.41**	-	**72.06**	-	**59.92**	-	-
2	**21**	**AB**	55.61	19.00	66.84	16.36									**61.23**	**17.68**	✓
3		**AE**	45.69	***-2.23***							68.41	0.00			**57.05**	***-1.12***	
4		**AF**	48.30	3.35									72.06	0.00	**60.18**	**1.68**	✓
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	
15		**DF**							36.55	***-10.83***			72.06	0.00	**54.31**	***-5.41***	
16		**EF**									71.54	4.58	74.67	3.62	**73.11**	**4.10**	✓
17	**22**	**ABC**	68.41	46.37	67.62	17.73	83.81	13.43							**73.28**	**25.84**	✓
18		**ABD**	73.37	56.98	66.32	15.45			50.91	24.20					**63.53**	**32.21**	✓
19		**ABE**	56.40	20.67	68.41	19.09					68.41	0.00			**64.40**	**13.25**	✓
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	
35		**CEF**					72.85	***-1.41***			72.32	5.73	74.93	3.99	**73.37**	**2.77**	
36		**DEF**							33.94	***-17.20***	71.02	3.82	72.85	1.09	**59.27**	***-4.10***	
37	**23**	**ABCD**	70.50	50.84	66.58	15.91	83.55	13.07	52.22	27.39					**68.21**	**26.80**	✓
38		**ABDE**	71.80	53.63	67.62	17.73			50.91	24.20	67.89	***-0.76***			**64.56**	**23.70**	
39		**ABDF**	70.76	51.40	65.27	13.64			54.05	31.85		0.00	72.06	0.00	**65.54**	**19.38**	✓
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	
49		**BCEF**			56.14	***-2.27***	71.02	***-3.89***			72.32	5.73	74.15	2.90	**68.41**	**0.62**	
51		**CDEF**					77.81	5.30	40.99	0.00	72.32	5.73	73.37	1.81	**66.12**	**3.21**	✓
52	**24**	**ABCDE**	69.97	49.72	67.10	16.82	84.33	14.13	50.39	22.93	68.41	0.00			**68.04**	**20.72**	✓
53		**ABCDF**	69.97	49.72	66.06	15.00	84.33	14.13	52.22	27.39			72.06	0.00	**68.93**	**21.25**	✓
54		**ABCEF**	68.93	47.49	69.71	21.36	84.33	14.13			71.80	4.96	74.67	3.62	**73.89**	**18.31**	✓
55		**ABDEF**	71.02	51.96	67.62	17.73			54.31	32.48	71.80	4.96	73.89	2.54	**67.73**	**21.93**	✓
56		**ACDEF**	68.67	46.93			84.60	14.49	49.35	20.38	71.80	4.96	75.46	4.71	**69.97**	**18.29**	✓
57		**BCDEF**			59.01	2.73	76.76	3.89	57.18	39.49	72.32	5.73	73.37	1.81	**67.73**	**10.73**	✓
58	**25**	**ABCDEF**	70.76	70.76	67.62	67.62	84.33	84.33	53.00	53.00	71.80	71.80	74.15	74.15	**70.28**	**70.28**	✓

Table 3 presented the evaluation results of feature adjustment method which are divided into two parts as follows:

1) The results of the accuracy performance of each model or each combination: The outcomes show the consistency between accuracy performance and a number of features, i.e., more number of features can get more accuracy performance, except in some cases.

2) The results of the improvement of accuracy: This work selects only classification model which gives the accuracy performance more than using 20 features set; 20 features set is the lowest number of features and it is based on only one feature of a user's profile.

The selection results get only 36 high performance classification models from 58 combinations which are marked by correct symbol in Table 3. The 36 models are obtained the accuracy performance more than using 20 features set in all classifiers. The selected models consist of 7 models from 21 features set (AB, AC, AD, AF, BD, CD, EF), 11 models from 22 features set (ABC, ABD, ABE, ABF, ACD, ACF, ADE, ADF, BCD, BDF, CDF), 11 models from 23 features set (ABCD, ABCE, ABCF, ABDF, ABEF, ACDF, ACEF, ADEF, BCDF, BDEF, CDEF), 6 models from 24 features set (ABCDE, ABCDF, ABCEF, ABDEF, ACDEF, BCDEF) and 1 models from 25 features set (ABCDEF).

5. Conclusion and Future Work

The development of recommendation rules set for PMRS by using the feature adjustment method makes more recommendation rules in the repository for appropriate suggestion. Moreover, the combinations of each feature support the incomplete registration function problem, i.e. if users fill blank in some registration function, PMRS can select recommendation rules which accord to specific data. Additionally, the method of this research will be applied to the development of a new personalized movie recommendation system in the near future.

6. References

[1] Li, P. and Yamada, S. A movie recommender system based on inductive learning. In Proceeding of *Cybernetics and Intelligent Systems*, *2004 IEEE Conference on*. IEEE, 2004. p. 318-323.

[2] Debnath, S., Ganguly, N., and Mitra, Feature weighting in content based recommendation system using social network analysis. In Proceedings of *the 17th international conference on World Wide Web*. ACM, 2008. p. 1041-1042.

[3] Mukherjee, R., Sajja, N., and Sen, S. A movie recommendation system–an application of voting theory in user modeling. *User Modeling and User-Adapted Interaction*, 2003, 13.1-2: 5-33.

[4] Zhang, Y. and Zhao, Y. (2003). Classification in multidimensional parameter space: Methods and examples. *Publications of the Astronomical Society of the Pacific*, 2003, 115.810: 1006.

[5] Quinlan, J.R. Induction of decision trees. *Machine learning*, 1986, 1.1: 81-106.

[6] Quinlan, J.R. *C4. 5: programs for machine learning*. Elsevier, 2014.

[7] Lin, C.F., Yeh, Y.-c., Hung, Y.H., and Chang, R.I. Data mining for providing a personalized learning path in creativity: An application of decision trees. *Computers & Education*, 2013, 68: 199-210.

[8] Wanapu, S., Chun C.C., Kajornrit, J., Niwattanakul, S., and Chamnongsri, N. Selecting feature grouping and decision tree to improve results from the learning object management model (LOMM). *Journal of Convergence Information Technology*, 2014, 9.3: 131.

[9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11.1 (2009), 10-18.