I.J. Intelligent Systems and Applications, **2013, 06, 53-60**
Published Online May 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2013.06.07

Prediction of Possible Business of a Newly Launched Film using Ordinal Values of Film-genres

Debaditya Barman, Research Fellow
Department of Computer Science and Engineering Jadavpur University, Kolkata - 700032, India
E-mail: debadityabarman@gmail.com

Rupesh Kumar Singha, MCA student
SikimManipal University, Kolkata - 700032, India
E-mail: rupeshsingha@gmail.com

Dr. **Nirmalya Chowdhury,** Associate Professor
Department of Computer Science and Engineering, Jadavpur University, Kolkata - 700032, India
E-mail: nirmalya_chowdhury@yahoo.com

Abstract— Film industry is the most important component of entertainment industry. Both profit and loss are very high for this business. Like every other business, business prediction system plays a vital role for this industry. Before release of a particular movie, if the Production Houses or distributors get any type of prediction that how the film will do business, then it will be very useful to reduce the risk of the investors. In this paper we have proposed a method using back propagation neural network for prediction about a given movie's profitability. Initially the entire range of profit-loss has been divided into a number of groups. The proposed algorithm can assign a given movie to it's appropriate profit-loss group. Note that, a similar such method has been successfully applied in the field of Stock Market Prediction, Weather Prediction and Image Processing.

Index Terms— Film Industry, Movie Genre, Profit Group, Artificial Neural Network, Back- Propagation Learning

I. Introduction

A movie [1], also called a film or motion picture, is a series of still or moving images. It is produced by recording photographic images with cameras, or by creating images using animation techniques or visual effects. The process of filmmaking has developed into an art form and has created an industry in itself.

Films are cultural artifacts created by specific cultures, which reflect those cultures, and, in turn, affect them. It is considered to be an important art form, a source of popular entertainment and a powerful method for educating or indoctrinating citizens. The visual elements of cinema give motion pictures a universal power of communication.

Film Industry is an important part of present-day mass media industry or entertainment industry (also informally known as show business or show biz). This industry [2] consists of the technological and commercial institutions of filmmaking: i.e. film production companies, film studios, cinematography, film production, screenwriting, pre-production, post production, film festivals, distribution; and actors, film directors and other film crew personnel.

The major business centers of film making are in the United States, India, Hong Kong and Nigeria. The average cost [3] of a world wide release of a Hollywood film or American film (including pre-production, film and post-production, but excluding distribution costs) is about $65 million. It can even go up to $300 million [4] (Pirates of the Caribbean: At World's End). Worldwide gross revenue [5] can be almost $2.8 billion (Avatar). Profit/loss is found to vary from a profit [6] of 2975.63 % (City Island) to a loss [7] of 1299.7 % (Zyzzyx Road). So it will be very useful if we can develop a prediction system which can predict about a film's business potential.

The initial task of profit-loss prediction in the movie business, is to predict, whether a movie will be profitable or not. Successful prediction about a movie's future can be a good lead for Distributors, Producers and Filmmakers. Their investment risk will be reduced abruptly. Many artificial neural network based methods are used to design for successful Stock Market Prediction[8], Weather Prediction [9], Image Processing [10], and Time Series Prediction [11], and Ambient Temperature Prediction system[12] etc.

We have divided the entire range of profit-loss range into a number of groups. Here we have proposed a method using back propagation neural network for prediction of this profit-loss group of a movie based on some pre-defined genres. Then we have compared the performance of our method with that of Naïve Bayes. Formulation of the problem is presented in the next section. Section III describes our proposed method. The method is presented in the form of an algorithm in section III-A. Experimental results on thirty two movies selected randomly from a given database can be found in section IV. Concluding remarks and scope for further work has been incorporated in section V.

II. Statement of the Problem

Every Film can be identified by certain film genres. In film theory, genre [13] refers to the method based on similarities in the narrative elements from which films are constructed. Most theories of film genre are borrowed from literary genre criticism. Some basic film genres are - action, adventure, animation, biography, comedy, crime, drama, family, fantasy, horror, mystery, romance, science-fiction, thriller, war etc. One film can belong to more than one genre. For instance, the movie titled "Pirates of the Caribbean: On Stranger Tides (2011)" belongs to [14] action, adventure, and fantasy genres.

Any film's success is highly dependent on its film genres. We can consider these genres as a film's attributes. We can then collect these attribute's data of past films. Based on these data we can predict about an upcoming film's future business.

We have used twenty attributes, which are basically twenty movie genres like action, adventure, animation, biography, comedy, crime, documentary, drama, family, fantasy, history, horror, musical, mystery, romance, science fiction, sport, thriller, war, and western. Note that the ordinal values of each attribute are used for experimentation. The movie genres are rated like following Table 1.

Table 1: Movie Genre Rating Chart

Rating Symbol	Meaning
A	Not present
B	Very Low
C	Low
D	Medium
E	High
F	Very High

We have divided the entire profit-loss range into 9 groups. We have listed them in the following Table 2.

Table 2: Movie's Profit Loss Rating

Rating Symbol	Meaning
A	A< - 49%
B	-49 ≤B<0 %
C	0 ≤C <50 %
D	50 ≤ D<150 %
E	150≤ E <250 %
F	250 ≤ F <350 %
G	350 ≤ G<450 %
H	450 ≤ H <550 %
I	I ≥550%

Artificial neural network [15] learning methods provide a robust approach to approximating real-valued, discrete-valued, and vector-valued target functions. For certain types of problems, such as learning to interpret complex real-world sensor data, artificial neural networks are among the most effective learning methods currently known.

Artificial neural network are applied in image recognition and classification, image processing, feature extraction from satellite images, cash forecasting for a branch of a bank, stock market prediction, decision making, temperature forecasting, atomic mass prediction, prediction of Thrombo-embolic Stroke, time series prediction, forecasting groundwater level.

Back-propagation is a common method of teaching artificial neural networks about how to perform a given task. It is a supervised learning method. It is most useful for feed-forward networks [16].

Back-propagation neural network is successfully applied in image compression [17], satellite image classification [18], irregular shapes classification [19], email classification [20], time series prediction [21], bankruptcy prediction [22], and weather forecasting [23].

III. Proposed Method

In this paper, we have proposed a method that uses a multilayer feed-forward neural network as shown in Fig.1. Note that, a multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.

Here the back-propagation algorithm [24] performs learning on the said multilayer feed-forward neural network. It iteratively learns a set of weights for prediction of the profit-loss group label (A/B/C/D/E/F/G/H/I) of instances.

As stated above, we have used twenty attributes (excluding the input for actual target output).Since we have used 20 genres, each genre having 6 ratings, we need 120 (20*6) nodes in the input layer. We have taken 18 nodes in the hidden layer and 9 nodes in the

output layer. Note that there is no strict guideline to choose the number of nodes in a hidden layer. However, usually the number of nodes in the hidden layer is to be between the input layer size and the output layer size. We have experimented using different number of nodes in the hidden layer but we have obtained consistently good result in respect of learning speed with eighteen nodes in the hidden layer.

Note that input nodes have received Boolean values that represent presence or absence of an individual genre rating. Note that for any genre it would have specific ordinal value out of six possible one. So the input feature vector can be represented as shown below.

$$\{x_{11}, x_{12}, \dots, x_{16}, x_{21}, \dots, x_{26}, \dots, x_{ij}\},$$

where $x_{ij} \in \{0,1\}$, x_{ij} is the value for individual genre rating. If the genre rating is present, then it will be 1, otherwise 0.

Boolean values are generated at the output nodes and only one output node will have two (that are Boolean 1/0) value. The Boolean output for all the output nodes taken together denotes the predicted profit-loss of a given input movie. For instance, if the output vector is [000000001] then it indicates that the input movie has been predicted to make a profit of equal or more than 550%.

We have used a database of released movie in the year 2011, 2010 and 2009 for training of the said network. After the network has been successfully trained it can be used for prediction of profit-loss group of new movie to be released. Our movie database contains 395 movies released from 2009 to 2011. In our experiment; we have selected 32 movies randomly from our dataset to make the test set. The rest 363 movies are used to create the training set. This test set is used for evaluating the efficiency of the trained back-propagation network. The experimental results are presented at the section IV.

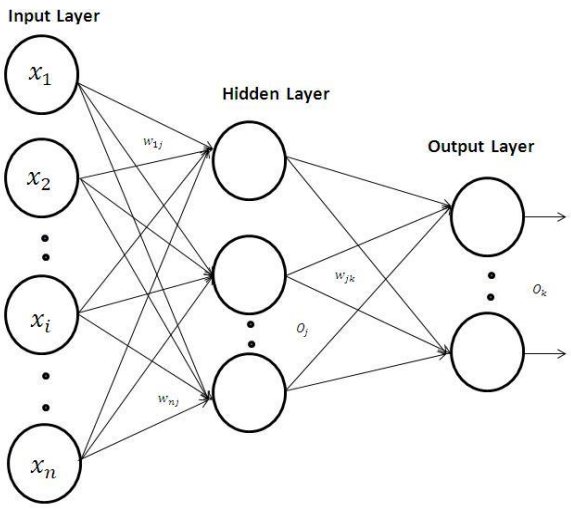

Fig. 1: A multilayer feed-forward neural network

Algorithm

Back-propagation Neural network learning for prediction, using the back-propagation algorithm.

Input:

- *D*, a data set consisting of the attributes of the movies and their profit/loss (profit/loss in Boolean value);
- *l*, the learning rate;

Output:

A trained neural network, which can predict profit loss group.

Method:

Step1. Initialize all weights and biases in network;

Step2. While terminating condition is not satisfied {

Step3. For each training instance ***X*** *in D {*

// propagate the inputs forward:

Step4. For each input layer unit j {

Step5. $O_j = I_j$ *// output of an input unit is its actual input value*

Step6. For each hidden or output layer unit j {

Step7. $I_j = \sum_i W_{i,j} O_i + \theta_j$*;//compute the net input of unit j with respect to the Previous layer, i*

Step8. $O_j = \frac{1}{1+e^{-I_j}}$*;} // compute the output of each unit j*

// Back propagate the errors:

Step9. For each unit j in the output layer

Step10. $Err_j = O_j(1 - O_j)(T_j - O_j)$ *; // compute the error*

Step11. For each unit j in the hidden layers, from the last to the first hidden layer

Step12. $Err_j = O_j(1 - O_j)\sum_k Err_k w_{jk}$ *; // compute the error with respect to the next higher Layer, k*

Step13. For each weight w_{ij} *in network {*

Step14. $\Delta w_{ij} = (l) Err_j \ O_i$ *; // weight increment*

Step15. $w_{ij} = w_{ij} + \Delta w_{ij}$ *}; // weight update*

Step16. For each bias θ_j *in network {*

Step17. $\Delta\theta_j = (l) Err_j$ *; // bias increment*

Step18. $\theta_j = \theta_j + \Delta\theta_j$ *; // bias update*

Step19. }}

Step20. Stop

At first, the weights in the network are initialized to small random numbers, bias associated with each unit also initialized to small random numbers.

The training instance from movie database is fed to the input layer. Next, the net input and output of each unit in the hidden and output layers are computed. A hidden layer or output layer unit is shown in Fig.2.

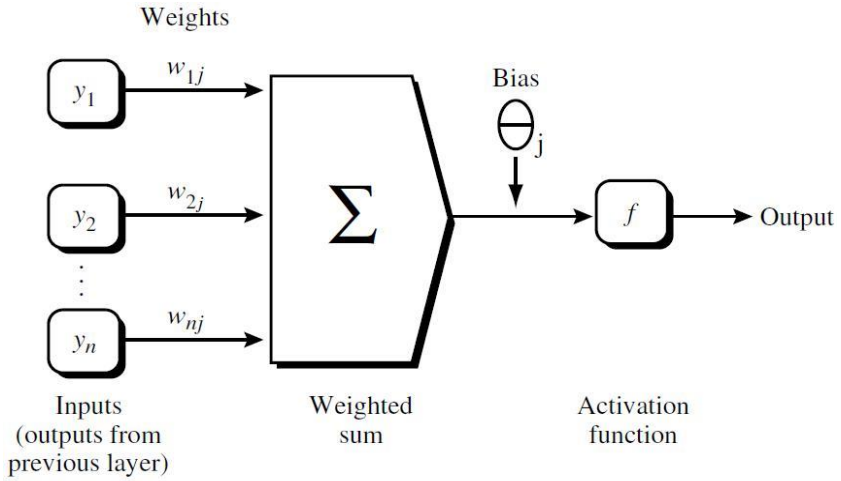

Fig. 2: A hidden or output layer unit j

The net input to unit j is

$$I_j = \sum_i w_{ij} O_i + \theta_j \quad (1)$$

Where w_{ij} is the connection weight from unit i, in the previous layer to unit j; O_i is the output of unit i from the previous layer; and θ_j is the bias of the unit.

As shown in the Fig.2, each unit in the hidden and output layers takes its net input and then applies an activation function to it. The function (sigmoid) symbolizes the activation of the neuron represent by the unit. Given the net input I_j to unit j, then O_j, the output of unit j computed as

$$O_j = \frac{1}{1+e^{-I_j}} \quad (2)$$

The error of each unit is computed and propagated backward. For a unit j in the output layer the error Err_j is computed by

$$Err_j = O_j(1 - O_j)(T_j - O_j) \quad (3)$$

Where, O_j is the actual output of unit j, and T_j is the known target value of the given training instance. The error of a hidden layer unit j is

$$Err_j = O_j(1 - O_j)\sum_k Err_k w_{jk} \quad (4)$$

Where, w_{jk} is the weight of the connection from unit j to a unit k in the next higher layer, and Err_k is the error of unit k.

The weights and biases are updated to reflect the propagated errors. Weights are updated by the following equations, where Δw_{ij} is the change in weight w_{ij}

$$\Delta w_{ij} = (l) Err_j \; O_i \quad (5)$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \quad (6)$$

l is the learning rate. In our experiment it is 0.1.

Biases are also updated, if $\Delta \theta_j$ is the change in θ_j then

$$\Delta \theta_j = (l) Err_j \quad (7)$$

$$\theta_j = \theta_j + \Delta \theta_j \quad (8)$$

The weights and biases are updated each time after all of the instances in the training set have been presented. In our experiment, we have used this strategy called *epoch updating*, where one iteration through the training set is called an epoch.

The training stops when

- All Δw_{ij} in the previous epoch were so small as to be below some specified threshold,

Or

- The percentage of instance misclassified in the previous epoch is below some threshold,

Or

- A pre specified number of epochs have expired.

In our experiment we have specified the number of epochs as 1000.

IV. Experimental Result

We have used the data of 395 movies released from 2009 to 2011. We have used ordinal values of twenty movie genres of each film as input, where the movie genres are action, adventure, animation, biography, comedy, crime, documentary, drama, family, fantasy, history, horror, musical, mystery, romance, science fiction, sport, thriller, war and western. The values of all attributes are rated as shown in Table 1. The database looks like as shown in Table. 3.

Table 3: Sample Of Data Used For Profit-Loss Prediction

Action	Adventure	Fantasy	.	.	.	Western	Profit
D	B	A	.	.	.	A	H
C	A	A	.	.	.	D	G
A	A	C	.	.	.	E	E
C	D	A	.	.	.	A	F

We have represented the profit group distribution in pie chart, Fig. 3.

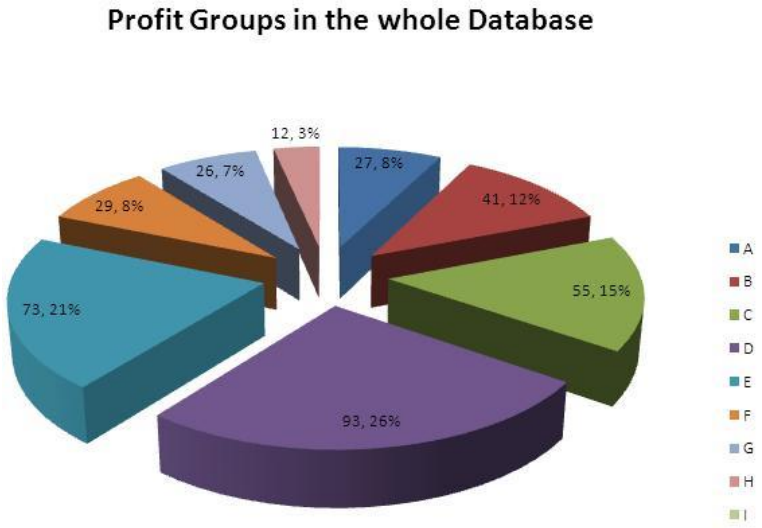

Fig. 3: Profit-loss distribution chart

We have selected 32 movies randomly from our dataset to create the test set. The rest 363 movies are used to create training set. We have carried out our experiment with WEKA 3.6.6 package. In this experiment we have used its inbuilt Multilayer Perceptron package. The "knowledge flow process" of our experiment using the package WEKA has been depicted in the Fig. 4. The multilayer perceptron used for this experiment is shown in Fig. 5.

byJiawei Han and MichelineKamber

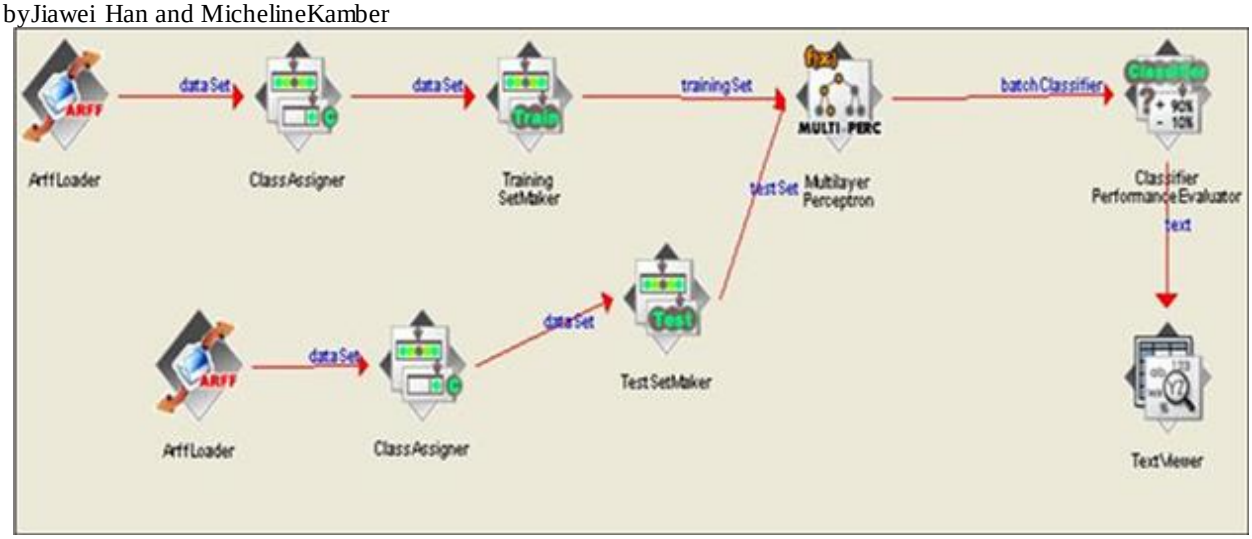

Fig. 4: MLP used for the experiment

We have given the details of prediction results for the movies of the test set in Table 4.

The success rate of 81.25% is good, if we consider the extreme uncertainty of the movie business. But in real life situations, prediction of profit-loss group is not good enough. For instance, for the movies released in USA in the year 2009, the loss percentage varied from −2.24338% to −97.2361%, whereas the profit percentage varied from 0.086793% to 1319.422%. Thus a system which can predict more accurately profit or loss of a new movie to be launched is highly necessary in movie industry. The main constrain to build such a system is the unavailability of continuous values of the movie genres till date.

We feel that a success rate of 81.25% provided by the proposed method is significant considering the highly unpredictable nature of the movie business world.

A summary of the experimental result with our proposed method and that of with Naïve Bayes classifier are shown in Table 5.

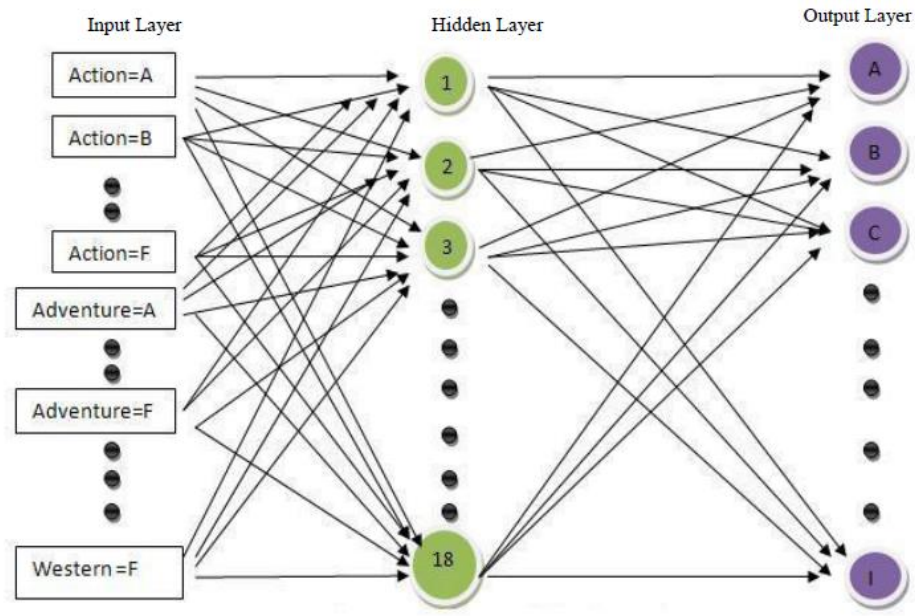

Fig. 5: Knowledge Flow layouts

Table 4: Experimental Result

Movie No.	Actual Profit-loss Group	Predicted Profit-loss Group	Classification	% of correct classification
1	A	B	Incorrect	81.25
2	A	A	Correct	
3	B	B	Correct	
4	C	C	Correct	
5	C	D	Incorrect	
6	C	C	Correct	
7	D	D	Correct	
8	D	D	Correct	
9	D	E	Incorrect	
10	D	D	Correct	
11	D	D	Correct	
12	I	I	Correct	
13	I	I	Correct	
14	D	D	Correct	
15	D	D	Correct	
16	D	D	Correct	
17	D	E	Incorrect	
18	D	D	Correct	
19	I	F	Incorrect	
20	I	I	Correct	
21	I	I	Correct	
22	H	H	Correct	
23	E	E	Correct	
24	E	E	Correct	
25	E	E	Correct	
26	E	E	Correct	
27	C	C	Correct	
28	C	C	Correct	
29	C	C	Correct	
30	C	C	Correct	
31	C	C	Correct	
32	C	B	Incorrect	

Table 5: Result Summary

Evaluation on test set		
Summary		
Measure	**Multilayer Perceptron**	**Naïve Bayes**
Correctly Classified Instances	26(81.25%)	28(87.5%)
Incorrectly Classified Instances	6(18.75%)	4(12.5%)
Kappa statistic	0.7661	0.8207
Mean absolute error	0.0622	0.1246
Root mean squared error	0.1824	0.2251
Total Number of Instances	32	32

As you can see from Table 5, the percentage both of correctly as well as incorrectly classified instances are available. In our experiment, only 26 out of a total of 32 instances have been correctly classified, leading to an accuracy of 81.25% achieved by MLP using ordinal value of genres. Note that only 28 out of a total of 32 instances have been correctly classified by Naïve-Bayes, leading to accuracy of 87.5%.

It may be noted that the value for Kappa statistic is a measure of quality of classification. It is considered to be a more robust measure than simple percent agreement calculation, since it takes into account the agreement occurring by chance. The value for Kappa coefficient for our proposed method using MLP is 0.7661 and .that of Naïve-Bayes is 0.8207. Mean absolute error indicates how close our predictions are to the target output. Root mean squared error gives a good measure of the model's accuracy.

V. Conclusion and Scope for Further Work

Note that, it is very difficult even for the human domain expert to predict the possible profit or loss of a new movie to be released. It seems that the genres of the movie play a significant role in making profit/loss of the movie, but it is very difficult to analytically establish the relation of the values of the genres of a given movie with the profit that it makes. In this paper we have attempted to develop a heuristic method using back-propagation neural network to solve this problem.

Further research work can be conducted in the search for more genres and/or division of an existing genre into subgenres that may led to a higher success rate of prediction.

Acknowledgements

This paper is an outcome of the work carried out for the project titled "In search of suitable methods for Clustering and Data mining" under "Mobile Computing and Innovative Applications programme" under the UGC funded "University with potential for Excellence – Phase II " scheme of Jadavpur University.

References

[1] en.wikipedia.org/wiki/Film.

[2] en.wikipedia.org/wiki/Film_industry

[3] www.the-numbers.com/glossary.php

[4] en.wikipedia.org/wiki/List_of_most_expensive_films

[5] en.wikipedia.org/wiki/List_of_highest-grossing_films

[6] en.wikipedia.org/wiki/City_Island_%28film%29

[7] en.wikipedia.org/wiki/Zyzzyx_Road

[8] "Stock Market Prediction with Back propagation Networks" by Bernd Freisleben Published in: IEA/AIE '92 Proceedings of the 5th international conference on Industrial and engineering applications of artificial intelligence and expert systems

[9] "Training back propagation neural networks with genetic algorithm for weather forecasting" by Gill, J.; Singh, B.; Singh, S.; This paper appears in: Intelligent Systems and Informatics (SISY), 2010 8th International Symposium

[10] "Multispectral image-processing with a three-layer back propagation network" by McClellan, G.E.; DeWitt, R.N.; Hemmer, T.H.; Matheson, L.N.; Moe, G.O.; Pacific-Sierra Res. Corp., Arlington, VA This paper appears in: Neural Networks, 1989. IJCNN., International Joint Conference

[11] Time Series Prediction and Neural Networks R.J.Frank, N.Davey, S.P.Hunt Department of Computer Science, University of Hertfordshire, Hatfield, UK.

[12] An Efficient Weather Forecasting System using Artificial Neural Network Dr. S. SanthoshBaboo and I.KadarShereef

[13] http://en.wikipedia.org/wiki/Film_genre

[14] http://www.imdb.com/title/tt1298650/

[15] "Machine Learning" by Tom Mitchell page 81

[16] http://en.wikipedia.org/wiki/Backpropagation

[17] "Image Compression with Back-Propagation Neural Network using Cumulative Distribution Function" by S. Anna Durai, and E. Anna Saro

[18] "Satellite Image Classification using the Back Propagation Algorithm of Artificial Neural Network." by Mrs. Ashwini T. Sapkal, Mr. ChandraprakashBokhare and Mr. N. Z. Tarapore

[19] "Irregular shapes classification by back-propagation neural networks " by Shih-Wei Lin, Shuo-Yan Chou and Shih-Chieh Chen

[20] "Email Classification Using Back Propagation Technique " by TaiwoAyodele, Shikun Zhou, RinatKhusainov

[21] "Parallel back-propagation for the prediction of time series" by Frank M. Thiesing, Ulrich Middelberg and Oliver Vornberger

[22] "Applying back propagation neural networks to bankruptcy prediction" by Yi-Chung Hu and Fang-Mei Tseng

[23] "An Efficient Weather Forecasting System using Artificial Neural Network" by Dr. S. SanthoshBaboo and I.KadarShereef

[24] Data Mining: Concepts and Techniques, 2nd ed.Page 328

Authors' Profiles

Debaditya Barman: Debaditya Barman is a Research Fellow in Computer Sc. and Engg. Department, Jadavpur University, India. He has obtained his B.E. and M.E. (Computer Sc. And Engg.)from Jadavpur University, India. His areas of interests are Machine Learning, Business intelligence, Data mining and Opinion mining.

Rupesh Kumar Singha: Rupesh kumar Singha did his graduation with honours from Calcutta University in 2010. At present he is pursuing MCA course of Sikim Manipal University. His fields of interests include Artificial intelligence and Data mining.

NirmalyaChowdhury: Nirmalya Chowdhury did his Ph.D. in Engineering from Jadavpur University, India in 1997. At present, he is working as Associate Professor in the department of Computer Science and Engineering, Jadavpur University, India. His fields of research include Pattern Recognition, Soft Computing, Natural Language Processing and Bioinformatics.

How to cite this paper: Debaditya Barman, Rupesh Kumar Singha, Nirmalya Chowdhury,"Prediction of Possible Business of a Newly Launched Film using Ordinal Values of Film-genres", International Journal of Intelligent Systems and Applications(IJISA), vol.5, no.6, pp.53-60, 2013.DOI: 10.5815/ijisa.2013.06.07

