International Journal of Hybrid Information Technology
Vol.9, No.2 (2016), pp. 417-426
http://dx.doi.org/10.14257/ijhit.2016.9.2.37

Research of Box-Office Prediction based on Rough Set and Support Vector Machine

Ling Liu[1] and Yang Zhao[2, 3]

[1]*School of Information Technology and Engineering Tianjin University of Technology and Education, Tianjin, 300222, China*
[2]*Department of Electronic and Information Technology, Jiangmen Polytechnic, Jiangmen, 529090, China*
[3]*College of Instrumentation Science and Electrical Engineering, Jilin University, Changchun, 130000, China*
E-mail: liulingtute@126.com

Abstract

In this paper, a novel prediction method for box-office is proposed based on the rough set data processing function and support vector machine (SVM) classification mechanism. The front-end processor, optimizes the input variables by attribute reduction, in order to improve the performance of classifier. Then, in view of the lack of guidance of scientific theory problem of domestic movie box office prediction, the classifier for the box-office prediction, the influence factors of the box office revenue as the input variables, the box-office income categories as output variables, data preprocessing and training test. Results show that the classifier can effectively solve the box office prediction problem, the results of the multilayer perceptron is better than that of Ramesh S. and Dursun D. using the prediction method, and the prediction error is less than 10%, to meet the requirements of the film market, show the powerful classification ability.

Keywords*: Rough Set, Support Vector Machine (SVM), box- office classification, box-office forecasting*

1. Introduction

As a product with a short lifetime, the film generates box office earnings during its schedule. Ticket sales is essential for risk evaluation of film investment and for the distributor [1-2]. Prediction of the ticket sales is not only an important means of ensuring the return on film investment and controlling film release risks, but also provides helpful information for investment decision [3]. Due to the large number of factors that influence ticket sales and the difficulty in quantifying these factors, it is really hard to accurately predict the profitability of films during their lifetime [4-5]. Ascertaining the factors in ticket sales and the interrelations among them is significant for reducing market risks and improving the operation and management of the film industry.

In [6], Marshall, *et al.,* proposed to estimate the total number of audience during the film's schedule using the historical film data. They also estimated the number of audience in the first week via the multiple linear regression algorithm, and applied the model devised by Sawhney, *et al.*, in [7] to estimate the aggregate number of audience in several weeks after the release. However, their method neglected the charm of the film for the audience, resulting in grave errors of estimation. Barman et al. proposed in [8] to predict the profitability of the film using the feedback neural networks. Despite the ability to provide accurate estimation locally, the structure of their proposed feedback neural network is too simple (containing only one hidden layer), and overlooks the effects from other important factors (*e.g.,* directors and actors), making it infeasible for practical

ISSN: 1738-9968 IJHIT
Copyright ⓒ 2016 SERSC

application. In [9], Sharda, *et al.*, proposed a multilayer neural network-based model which can classify film's ticket sales by integrating multiple film properties that influence ticket sales. And the classification accuracy is used as the major performance metric for the classification evaluation of the model, resulting in great classification effectiveness. But the interval used in the output layer of their classification model is too large (*e.g.*, the ticket sales interval of [100000, 1000000]), decreasing their model's feasibility.

To address existing problems with ticket sales prediction, we propose a hybrid classification method based on the rough set and the support vector machine. Considering random fluctuations of ticket sales, we improve the algorithm and the prediction process by specifying the range of ticket sales fluctuations. Our model can not only ensure accuracy of the final prediction result, but also provides insights into film investment risk control.

2. Basics of the Rough Set Theory

2.1. Classic Rough Set Theory

The rough set (RS) theory provides a tool for addressing the classification problem that involves ambiguous, inaccurate or incomplete information [10-11].

Given a limited non-empty set U (universe of discourse), and let R denote a set of equivalent relations on U, then the bivariate pair (U, R) forms an approximation space. The equivalent relations R partition U into mutually disjoint subsets $E_i, i=1,2,\cdots,n$. E_i and the empty set are called the basic set denoted by $U/R=\{E_1,E_2,\cdots,E_n\}$. Let X be a subset of U. If X cannot be accurately represented by the union of the basic sets, then X is called the rough set. The union of all basic sets contained in X is the lower approximation of X, denoted by $R_*(X)$. The union of basic sets whose intersection with X is not empty is the upper approximation of X, denoted by $R^*(X)$. It can be mathematically defined as:

$$\begin{cases} R_*(X)=\{x\in U\mid [x]_R\subseteq X\} \\ R^*(X)=\{x\in U\mid [x]_R\bigcap X\neq\Phi\} \end{cases} \tag{1}$$

where x is an object in U, $[x]_R$ denotes the equivalent class that is obtained by partitioning U with the equivalent relation R and it contains x.

Let U be the universe of discourse, P and Q be the two equivalent relations (i.e. knowledge) in U; The partition of P and Q in U is X and Y, respectively: $X=\{X_1,X_2,\cdots,X_n\}, Y=\{Y_1,Y_2,\cdots,Y_n\}$. Then, the P-positive domain of Q is denoted by $POS_P(Q)$ and defined as:

$$POS_P(Q)=\bigcup_{X\in U/Q} P_*(X) \tag{2}$$

The P-positive domain of Q refers to the set of objects that can be correctly partitioned into the equivalent class of Q via the information from U/P. The dependence degree of the knowledge is defined as:

$$k=\gamma_P(Q)=\left|POS_P(Q)\right|/\left|U\right| \tag{3}$$

where $|\bullet|$ denotes the cardinality of $\bullet$.

The RS theory classifies the knowledge via knowledge reduction. The properties of the information system are not equally important and it is possible that some properties are redundant. Therefore, knowledge reduction can be done to delete irrelevant or unimportant property knowledge without compromising the classification ability of the system, and to extract the properties and rules that can best represent system features and variations, resulting in the most simplified system. To determine the importance of some

 Copyright ⓒ 2016 SERSC

properties, the strategy adopted in this paper is to add the property one by one and study how the classes will change after this property is added. If the addition of this class lead to radical changes of classes, then it means that this property is intense, i.e. important. Otherwise, it means that this property is not intense, i.e. unimportant.

$$\sigma_{PQ}(p) = \gamma_{P+\{p\}}(Q) - \gamma_P(Q) \tag{4}$$

2.2. Fuzzy Rough Set

When the RS theory is used for data analysis, it requires all properties of the information system to be represented with discrete values. Although there are many methods for discretizing the continuous data, they neglect whether elements within the class are differentiable. Even if it can be known which set the element belongs to, it is unknown in which degree the element belongs to the set. From this perspective, this causes loss of information. The use of the fuzzy rough set enables the inter-class elements to be differentiated via the fuzzy membership function during property reduction, thus resulting in a reduction that incurs a slight loss of information.

By substituting the fuzzy set for the accurate set and introducing the fuzzy similarity relation to the universe of discourse as a replacement for the accurate similarity relation [12-13], the classic fuzzy set theory can be extended to the fuzzy rough set [13]. The fuzzy lower and upper approximation is defined as:

$$\begin{cases} \mu_{P_*X}(F_i) = \inf\limits_x \max\left\{1-\mu_{F_t}(x), \mu_X(x)\right\} & \forall i \\ \mu_{P_*X}(F_i) = \sup\limits_x \min\left\{\mu_{F_t}(x), \mu_X(x)\right\} & \forall i \end{cases} \tag{5}$$

where Fi denotes the fuzzy equivalent class that belongs to U/P, the bivariate pair <P*X, P*X>is called the fuzzy rough set.

For x belonging to the fuzzy positive region in U, it is defined as [10]:

$$\mu_{POS_P(Q)}(x) = \sup_{X \in U/Q} \mu_{P_*(X)}(x) \tag{6}$$

Correspondingly, the dependence degree is defined as:

$$\gamma'_P(Q) = \left|\mu_{POS_P(Q)}(x)\right| / (U) = \sum_{X \in U} \mu_{POS_P(Q)}(x) / |U| \tag{7}$$

3. Basics Principle of SVM

Support vector machine (SVM) is a wholly novel machine learning algorithm proposed by V. Vapnik, *et al.*, from the Bell laboratory who have studied statistical learning for over 30 years. By combining the maximum margin classifier and the kernel-based method, SVM exhibits great generalization ability [14]. Its basic idea is as follows: transform the input space into a high-dimension space via the nonlinear mapping j, and then find the nonlinear relation between the input variable and the output variable in the high-dimension space. The algorithm only uses the inner product in the high-dimension space. So by introducing proper kernel functions, the inner product calculation in the high-dimension space can be done with the functions in the original space. And linear classification based on a nonlinear transformation can also be achieved [15].

The core of the SVM idea is to map the data x into the high-dimension feature space via a nonlinear mapping j before performing linear regression in this space. Generally, the regression problem can be described as follows: for a given training sample, the learning machine learns the relation between the input variable and the output variable. Consider the given training data $\{(x_i, y_i), i = 1, 2, \cdots, l\}$, where, $x_i \in R_n$ is the n-dimension input value of the i[th] sampling point, $y_i \in R$ is the corresponding target value. l is the number

of training samples. The objective is to find a function $f(x)$, which can remarkably approximate to all sample points [6]. Generally, the estimation function of SVM can be given by:

$$f(x) = <\omega^T j(x)> + b \tag{8}$$

where $f(x)$ denotes the regression function, w and b denote the normal vector and the deviation, $j(x)$ denotes the characteristic mapping function.

Then, the standard support vector regression algorithm can be formulated as the following problem:

$$\min \frac{1}{2}\omega^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^{\omega}\right) \tag{9}$$

$$s.t.\begin{cases} y_i - <\omega^T j(x_i)> - b \le \varepsilon + \xi_i \\ <\omega^T j(x_i)> + b - y_i \le \varepsilon + \varepsilon_i^{\varepsilon} \\ \xi_i, \xi_i^{\omega} \ge 0, i = 1, 2, \cdots, l \end{cases} \tag{10}$$

where C is the punitive coefficient, ξ_i is the relaxation variable, and ε is the loss function.

The Lagrange multiplier is used to find the solution. We introduce the Lagrange multiplier α_i, α_i^*, and the kernel function $K(x_i, x_j) = <j(x_i) \square j(x_j)>$ to solve the Lagrange function. Via derivation, its dual optimization problem is formulated as:

$$s.t.\begin{cases} \sum_{i=1}\left(\alpha_i - \alpha_i^*\right) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \qquad i = 1, 2, \cdots l \tag{11}$$

$$\begin{aligned} &\max W\left(\alpha_i, \alpha_i^*\right) = \frac{1}{2}\sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_i^*\right)K\left(x_i - x_j\right) \\ &-\varepsilon\sum_{i=1}^{l}\left(\alpha_i + \alpha_i^*\right) + \sum_{i=1}^{l} y_i\left(\alpha_i - \alpha_i^*\right) \end{aligned} \tag{12}$$

After the above quadratic optimization problem is solved, the general formula can be rewritten as:

$$\begin{aligned} f(x) &= \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right) < j(x_i)\, j(x) > + b \\ &= \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)K(x_i, x) > + b \end{aligned} \tag{13}$$

where b is computed via SVM, the kernel function $K(x_i, x)$ is the dot product that any symmetrical kernel function satisfying Mercer conditions corresponds to in the feature space. Common kernels include the linear kernel, polynomial kernel and the radial base function (RBF).

4. Prediction Model

Figure 1 shows the ticket sales classification and prediction model based on the rough set and SVM. The decision table is constructed after the input variables are chosen and regularized. Redundant properties and conflicting samples are deleted via property reduction without compromising the classification ability. Finally, the obtained data is used for the training and testing of SVM. Main processes are given below:

 Copyright ⓒ 2016 SERSC

(1) Input the training data and testing data. Use the cross validation strategy to partition the samples. In this way, we can make full use of existing samples to test the network performance and correctly represent the model's performance.

(2) Pre-process the data. Process the ticket sales data, collecting the data, deleting redundant data, outdated data and invalid data.

(3) Select the input variable and normalize it. Based on the characteristics of the film, extract the features regarding ticket sales as the input variables and normalize them.

(4) Construct the decision table. Define the ticket sales prediction factors as the conditional properties, and the ticket sales as the decision properties. Then, construct the decision table and discretize it.

(5) Reduce decision table properties. Reduce the table using the property reduction algorithm based on the importance of properties. Reduce the input variables without compromising the classification ability.

(6) Test and train the classification ability of SVM.

(7) Check whether the classification quality meets the requirement. If it does, then output the result. Otherwise, select the features anew.

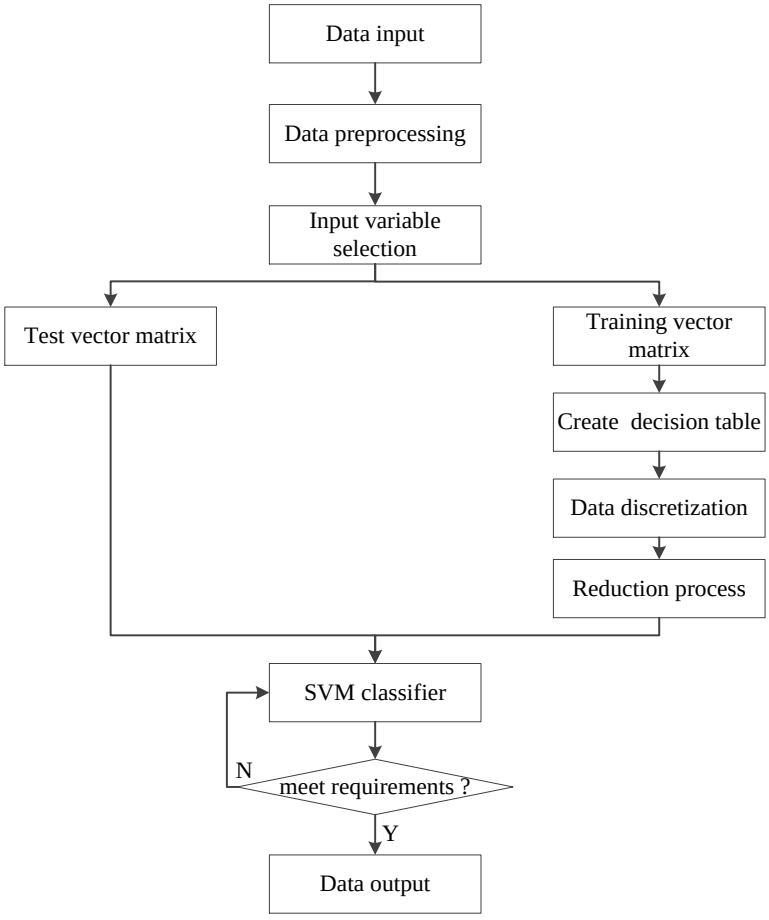

Figure 1. The Principle Diagram of the Forecasting Model

5. Prediction Results Analysis

5.1 Cross Validation Strategy

The BP algorithm designed for the multi-hidden-layer network is chosen as the training algorithm of the neural network classifier. The cross validation strategy is

used for the evaluation of the prediction performance of the neural network classifier. Here, the samples are uniformly partitioned into 6 groups. That is, we use the six fold cross validation approach. The sum of each column in this matrix denotes the actual number of films of the corresponding type, and the sum of each row denotes the number of films of the corresponding type estimated by the classifier. Hence, the element in the diagonal denotes the accurately estimated number of films of the corresponding type. And the type of films in each group tend to be uniformly distributed. The groups are shown in Table 1.

Table 1. Groups of 6-Fold Cross Validation

No.	Number of different kinds of movies						Sum
	1	2	3	4	5	6	
1	6	8	8	8	5	6	41
2	6	8	8	7	6	5	40
3	6	7	9	7	6	5	40
4	5	8	8	8	6	5	40
5	5	8	8	8	6	5	40
6	5	8	8	8	6	5	40
Sum	33	47	49	46	35	31	241

5.2 Performance Metrics of Ticket Sales Prediction

As in the traditional performance evaluation method, the percent accuracy (APHR=number of correctly classified samples/the total number of samples of this type) is used to measure the prediction quality of the neural network classifier [9]. There are two main performance metrics: absolute accuracy (Bingo), defined as the ratio of the number of films whose type is correctly predicted to the total number of films of the corresponding type; relative accuracy (1-Away), defined as the ratio of the number of films whose type and neighboring types are correctly predicted to the total number of films of the corresponding type. Both metrics can be computed as in Equations 14 and 15. These hit ratios indicate the average accuracy of the classification model's actual output with respect to the expected output.

$$Bingo = \frac{1}{n}\sum_{i=1}^{c} p_i \tag{14}$$

$$1-Away = \frac{1}{n}\sum_{i=1}^{c}\left(p_{i-1} + p_i + p_{i+1}\right) \tag{15}$$

where C denotes the total number of classes (=6), n denotes the total number of films of the ith type, pi denotes the number of films which are predicted to be the ith type. The films are partitioned into 6 classes, so when i<1 or i>6, we have pi=0.

5.3 Prediction Results and Performance Analysis

The previous studies demonstrate that the multi-layer perceptron (MLP) provides the greatest prediction performance. And it has been proved that its effectiveness is better than the traditional statistics-based classification approaches, such as the decision classification method, the regression classification method and the decision tree method. So our proposed method is directly compared with MLP.

 Copyright ⓒ 2016 SERSC

The reduced results are input to the SVM classifier for training and testing. RBF is chosen as the kernel function. Based on several experiments, the punitive factor C is set to 460 and $\sigma^2 = 5$. Training and classification simulations are performed via MATLAB 7.0. The mixed matrix usually used in the classification problem is adopted here for the representation of the classification results. The prediction results are given in Table 2.

Table 2. Prediction Result of Box-Office by Proposed Method

No.	1	2	3	4	5	6
1	14	12	5	0	0	0
2	11	22	13	3	1	0
3	7	11	19	14	0	1
4	1	2	11	25	8	1
5	0	0	1	5	19	6
6	0	0	0	0	6	23
Bingo /%	42.4	46.8	38.8	53.2	55.9	74.2
1-Way /%	75.8	95.7	87.8	93.6	97.1	93.5

Simulation results of MLP are shown in Table 3.

Table 3. Groups of 6-Fold Cross Validation by MLP Method

No.	1	2	3	4	5	6
1	10	3	1	1	0	0
2	11	15	11	3	3	0
3	3	17	16	2	4	0
4	6	7	13	20	8	4
5	1	5	4	13	15	9
6	2	0	4	7	5	8
Bingo /%	30.3	31.9	32.7	43.5	42.9	58.1
1-Way /%	63.6	74.5	81.6	76.1	80	87.1

By comparing the result achieved with the input variables that are not reduced with the result achieved with the reduced input variables, it can be found that due to property reduction, the absolute accuracy of the classifier improves by 10.5% in average, and its relative accuracy improves by 4.8% in average. Comparison of both the individual group's prediction results and the overall results between SVM and MLP demonstrates that the proposed method outperforms MLP.

Copyright ⓒ 2016 SERSC

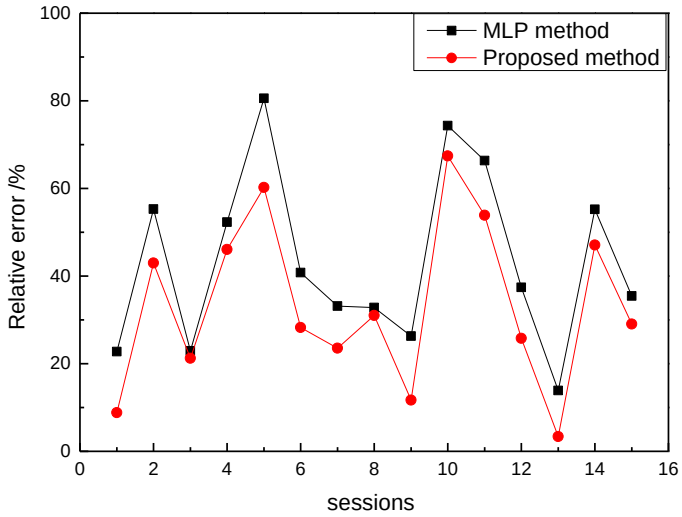

Figure 2. The Box-Office Prediction Relative Error Comparison in the Second Week

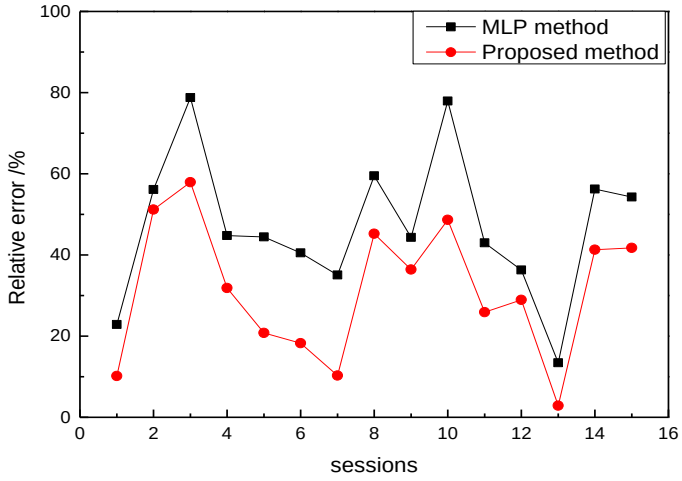

Figure 3. The box-Office Prediction Relative Error Comparison in the Third Week

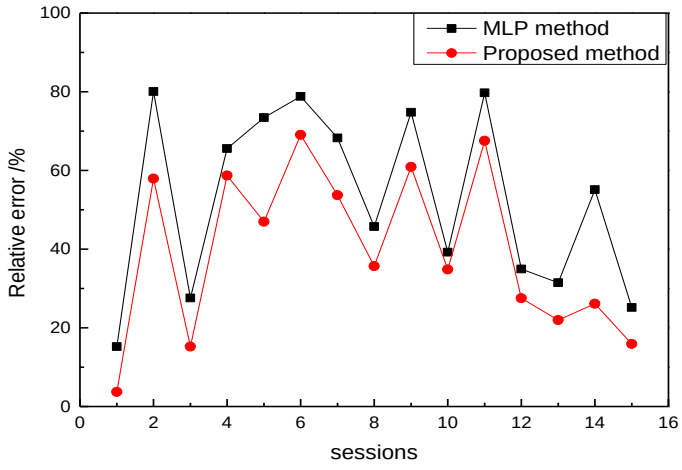

Figure 4. The Box-Office Prediction Relative Error Comparison in the Fourth Week

Copyright ⓒ 2016 SERSC

Figure 2-Figure 4 show the comparison of box-office prediction results between traditional MLP method and the proposed forecasting method in this paper in the second week to the fourth week. From the figures, it can be clearly seen that, the prediction precision of prosed method based on rough set and SVM is superior to the MLP method. The average relative error are down 9.93%, 15.7% and 13.3% from second week to the fourth week, respectively.

5. Conclusion

In this paper, we proposed a novel box-office forecasting method by combining rough sets attribute reduction and support vector machine classification. In the selection of input variables and to determine the initial value, using the statistical method, the value of more scientific and reasonable. Results show that its accuracy is higher than that of MLP by 10%. So our proposed method is effective and can meet the industry's requirements. This model can be used for film investment and producers released in a film before filming even forecast the output, but also can provide a reference for the company to choose the film, help to make scientific and rational decision-making.

However, this paper is just a theoretical model and presents simulation experimental result. In the future, we consider combining with the requirements of the filmdom, developing an easy operation human-machine interface application program. This model can be put into practical use, in order to enhance the accuracy of prediction of the box-office and create economic benefits

References

[1] H.-T. Thorsten, B.-H. Mark and H. Torsten, "Conceptualizing and Measuring the Monetary Value of Brand Extensions: The Case of Motion Pictures", Journal of Marketing, vol. 73, no. 6, **(2009)** November, pp. 167-183.
[2] M. Sangkil, K.-B. Paul and I. Dawn, "Dynamic Effects among Movie Ratings, Movie Revenues, and Viewer Satisfaction", Journal of Marketing, vol. 74, no. 1, **(2010)** January, pp. 108-121.
[3] E. Jehoshua, E. Anita and A.-L. Mark, "The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions", Marketing Science, vol. 25, no. 6, **(2006)** November, pp. 638-661.
[4] B.-H. Chang and E.-J. K, "Devising a Practical Model for Predicting Theatrical Movie Success: Focusing on the Experience Good Property" Journal of Media Economics, vol. 18, no. 4, **(2005)** November, pp. 247-269.
[5] E. Montañés, S.-V Ana and R.-Q. José, "Ordinal classification/regression for analyzing the influence of superstars on spectators in cinema marketing", Expert Systems with Applications, vol. 41, no. 18, **(2014)** December, pp. 8101-8111.
[6] P. Marshall, M. Dockendorff and S. Ibáñez, "A forecasting system for movie attendance", Journal of Business Research, vol. 66, no. 13, **(2013)** October, pp. 1800-1806.
[7] M. Sawhney and J. Eliashberg "A parsimonious model for forecasting gross box-office revenues of motion picture" Marketing Science, vol.15, no. 2, **(1996)**, pp. 113-131.
[8] D. Barman, Y.-N. Chowdhur and R.-K. Singha, "To predict possible profit/loss of a movie to be launched using MLP with back- propagation learning", in Proceedings of the 2012 International Conference on Communications, **(2013)**, Piscataway, USA.
[9] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks", Expert Systems with Applications, vol. 30, no. 2, **(2006)** February, pp. 243-254.
[10] C.-Y. Wang and B.-Q. Hu, "Fuzzy rough sets based on generalized residuated lattices", Information Sciences, vol. 248, no. 1, **(2013)** November, pp. 31-49.
[11] X.-P. Kang, D.-Y. Li, S.-G. Wang and K.-S. Qu, "Rough set model based on formal concept analysis", Information Sciences, vol. 222, no. 10, **(2013)** February, pp. 611-625.
[12] A.-M. Radzikowska and E.-E. Kerre, "A comparative study of fuzzy rough sets", Fuzzy Sets and Systems, vol. 126, no. 2, **(2002)** March, pp. 137-155.
[13] D. Dubois and H. Prade, "Rough Fuzzy Sets and Fuzzy Rough Sets*", International Journal of General Systems, vol. 17, no. 2-3, **(1990)** April, pp. 191-209.
[14] C. Nello and S.-T. John, "Introduction to support vector machines", Beijing: Mechanical Industry Press, **(2005)**, pp. 93-160.

Copyright ⓒ 2016 SERSC

[15] M.-H. He, L. Lu and S.-h. Liu, "Forecasting regional logistics amount based on fuzzy-rough set and SVM", Journal of Transportation Systems Engineering and Information Technology, vol. 12, no. 3, **(2012)** March, pp. 129-134.

 Copyright ⓒ 2016 SERSC