Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220041

Strange Approach of Movie Rating Prediction Using Logistic Regression Comparing to Gaussian Naive Bayes Algorithm

Ganesh R[a, 1] and Kalaiarasi S [b]
[a]*Research Scholar, Department of CSE, Saveetha School of Engineering,*
[b]*Asst. Professor, Department of AI & DS, Saveetha School of Engineering,*
[a,b] *SIMATS, Chennai, India*

Abstract. The aim is to find movie ratings using logistic regression and comparing the result with naive bayes based on Accuracy. A total of 6040 samples were collected from movie datasets available in kaggle. Two algorithms are used; one is Logistic Regression and another is naive bayes algorithm. The computation processes were executed and verified for exactness. Sample size N=5 is taken for both algorithms. SPSS was used for predicting significance value of the dataset considering G-Power value as 80%. Logistic Regression achieved mean accuracy of 80.83% when compared to Naive Bayes Algorithm with 82.53%. Results were obtained with a level of significance with 0.003 ($p<0.05$). Applied strange recommendation model confirms to have higher accuracy than Naive Bayes algorithm.

Keywords. Naive Bayes Algorithm, Novel Logistic Regression, Normalization, Machine Learning, Artificial Intelligence.

1. Introduction

In today's hectic world, movie rating systems are becoming increasingly important. People are always in state of mind to complete all their activities in a 24-hour period. In their busy life style they are not able to spend time on their personal chaos like going outing with family, watching movie, etc. They don't want to waste their time in unworthy things and they are not able to make decision on right thing to be done. As a result many artificial intelligence based recommendation systems are available. Among them movie rating systems are significant since they assist them in making informed decisions without having to use their cognitive resources. The goal of a movie rating system is to find stuff that would be attractive to a specific person. Movie rating systems are Artificial Intelligence-based algorithms that scan all available possibilities and generate a personalized list of stuff that are interesting and relevant to a specific person. These recommendations are based on their profile, browsing history, what other people with similar characteristics are watching, and your likelihood of watching those

[1]Ganesh R, Department of CSE, Saveetha School of Engineering, SIMATS, Chennai, India.
E-mail: kalaiarasis.sse@saveetha.com

films.Sentiment analysis paves a vital role in movie rating [1]. Netflix effects was given and also extended that Recommendations are not a new concept [2]. In the situation where e-commerce was not applied, even the sales person in retail stores recommend items to customers for upselling and cross-selling, in order to maximize profit[3]. Application of recommended system is to help users to find their items according to their interest, help item providers to provide their items to the appropriate user and to identify relevant products for every user. Applications such as BetaSeries, Cineast, and JustWatch, one can track down top movies of importance and even it can track upcoming movies [4].

There are 13 research articles published in IEEE Explore and around 5400 articles found in Google scholar. In recent times, surveys of machine learning algorithms for movie ratings were explored mostly as they predicted 80% of output accuracy. A movie suggestion framework dependent on collective separating approaches[5]. Later, reviews of AI calculations for film evaluations were investigated for the most part as they anticipated 80% yield precision. A film idea system reliant upon aggregate isolating optimization approaches[6-8]. This specific article is referred to multiple times by the customers. A structure has been presented that solidifies both shared and content-based techniques[9]. Additionally made an investigation of both the conventional proposal methods and this systems have certain setbacks, hence, another structure was proposed, which is a blend of Bayesian organization and shared[10]. The Efficient movie recommendation algorithm was implemented based upon improved *k*-cliques, this paper has more than 10 citations. The findings are performance results shows that these methods improve more accuracy in the movie recommendation system than other methods used in this experiment [11]. The aim is to improve the accuracy in movie rating.

2. Materials and Methods

This study was done in the Machine Learning Laboratory at Saveetha School of Engineering. This examination comprises two example bunches i.e, KNN Algorithm and novel Logistic Regression. Each gathering contains 392 examples with a pretest power of 80% taken for testing with alpha of 0.05. The dataset utilized for division was taken from the Kaggle dataset storehouse. There are various strategies associated with executing this proposal framework which incorporates different fields of Data Mining, Clustering and Bayesian Network procedure [12].

2.1.Logistic Regression

Logistic Regression is one of the important machine learning algorithms that mainly concentrates on classification example 0 or 1 and pass or fail. It uses sigmoid function for mapping the predicted values to probabilities and decides on which values to pass as output and not to pass based on the input parameters passed and helps in predicting categorical dependent variables using independent variables. The general equation for Logistic regression was shown below,

log[y/y-1] = b0+b1x1+b2x2+.....+bnxn
Where x1 ,x2,.....xn are the observations and predictive variables, log[y/y-1] is sigmoid function ,b0 is intercept and biare coefficients.

The primary attributes such as dataset score, id_student and date_submited, were used for predicting accuracy percentage of Learning Achievement Model using Logistic Regression Machine Learning Algorithm.

2.2. Naive Bayes

Naive Bayes is built on Bayes Theorem. It is a supervised machine learning algorithm and it is used for classification problems. SVM is a supervised machine learning algorithm and it can be used to perform Regression and Classification. The required packages were imported and the dataset was submitted into a code environment.Then next step is cleaning data in order to remove punctuation marks and other symbols.

It was done by Regex. After removing punctuation and symbols the data is stored in the same data frame. After cleaning the data, it is sent as input to perform training and to test prediction by using test data. Number of epochs is the amount your model will rotate and learn about, and size of the batch is the amount of data your model will see at the same time. As we are trained in small data sets on a few models, epochs will fit. Division of sum of data test labels by length of predicting labels will give accuracy.

2.3. Testing Setup

Dataset for testing and training was collected. Data preprocessing was completed. Data cleaning is done on the dataset, then concatenating and shuffling was done. Information set that contains actual data needed for classifier is converted. Split the dataset into training as 30% and testing set as 70%. Cross Validation needs to be done automatically, and split function generates and now implements Machine Learning Classifier using training dataset to train the classifier.

Once the training is completed, the classifier with the help of testing dataset checks the trained classifier to urge the anticipated accuracy got from the classifier. For training, the test set size = 30% of total dataset and training set = 70%. Whole dataset is fitting for training both Algorithms. Accuracies of both models were tested with different sample sizes from 50 to 1000.

3. Results

Table 1 shows that the dataset consists of 4 columns, visually, Column 1, indicates Serial number (S.No), Column 2, indicates, Internet Movie ID (IM_Id), Column 3, indicates, Rating range, which is from 0 to 5, and Column 4, indicates location of customer.

From Fig. 1, it was observed that this shows accurate ratings of movies. The following tables give us top 25 Movie ratings given to the customer to visualize himself and to make a proper choice of the best movie to enjoy. X-axis shows the number of customers who gave ratings, Y- axis shows ratings as 1 or 2 or 3 or 4 or 5.

Table 1. Movie Ratings of Customers, the dataset column 1 that indicates Rating range, which is from 0 to 5, and Column 2 indicates location of the customer.

Ratings	Location
5	97830076
3	978302109
3	978301968
4	97830027

In Table 2, it was observed that there is Performance comparison of algorithms with 5 iterations, Data collection from the N=5 samples of the dataset for Naive Bayes and the Logistic Regression algorithm with the highest accuracy of 86.23% and 87.98% in sample 5, using the training data and testing data, are respectively 70% and 30%.

Table 2. Performance comparison of algorithms with 5 iterations N=5 sample size of the dataset for Naive Bayes and Logistic Regression algorithm with the highest accuracy of 86.23% and 87.98% in sample 5, using the training data and testing data 70% and 30% respectively.

Naive Bayes Algorithm Accuracy %	Logistic Regression Algorithm Accuracy %
76.23	77.98
78.23	79.98
80.23	81.98
83.23	84.98
86.23	87.98

In Table 3, it was observed that,T-test comparison, Group Statistic analysis, representing Naive Bayes (mean accuracy 80.83%, SD 3.975) and novel Logistic Regression (mean accuracy 82.53% SD 3.975).

Table 3.T-test comparison, Group Statistic analysis representing Naive bayes (mean accuracy of 80.83%, SD=3.975) and Logistic Regression (mean accuracy of 82.53% SD=3.975)

Performance	Algorithm	N	Mean	SD	Error
Accuracy	Naive Bayes Algorithm	5	80.83	3.975	1.778
Accuracy	Logistic Regression-Algorithm	5	82.53	3.975	1.778

In Table 4, it was shown that Independent Samples Test, for Logistic Regression and Naive Bayes (mean difference -1.250 and SD error difference 2.514 with significance 2-tailed 0.003 and respectively). F-was Fisher test, which was applied for Testing of Hypothesis, T-test was applied for comparing two groups with 95% confidence interval and df was degrees of freedom for n samples.

Table 4. Independent Sample Test for SLogistic Regression and Naive Bayes (mean difference -1.250 and SD error difference 2.514) provides statistical significance of 0.003 (2-tailed)

	Variance	**F**	**Sig**	**t**	**df**	**Sig (2-tailed)**	**Mean difference**	**std error diff**	**lower bound**	**upper bound**
Accuracy	Equal variances assumed	0.000	0.03	-.696	8	0.003	-1.750	2.514	-7.547	-4.047
Accuracy	Equal variance not assumed			-.696	8.00	0.003	-1.750	2.514	-7.547	-4.047

From Figure 2, it was discovered that it compares the mean accuracy of the Logistic Regression algorithm to the Naive Bayes classifier. Mean accuracy of Logistic Regression is better than Naive bayes. And the standard deviation of Logistic Regression is slightly better than Naive bayes. X Axis: Logistic Regression vs Naive Bayes Algorithm Y Axis: Mean accuracy of detection ± 1 SD.

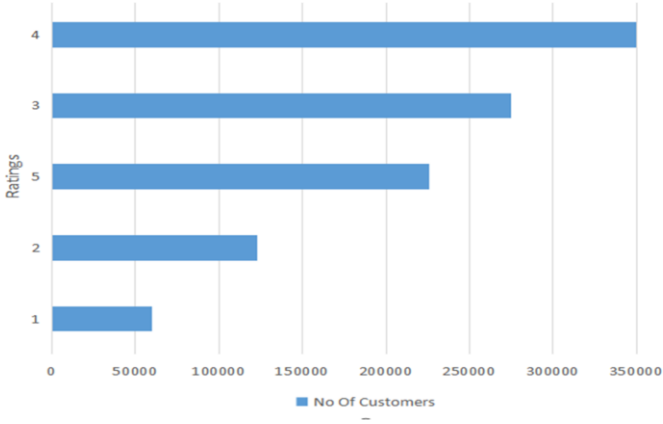

Figure 1. Top twenty five Movie Ratings given to the customer

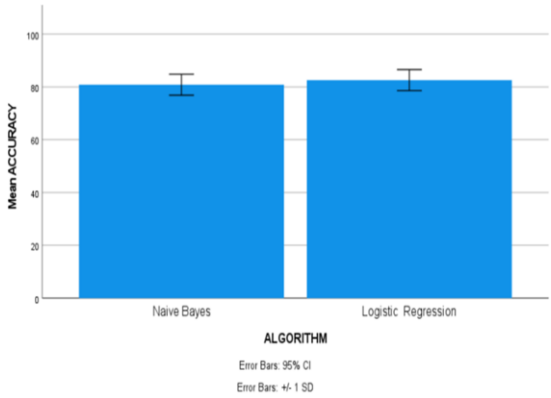

Figure 2. Bar graph of Logistic Regression algorithm and Naive bayes classifier in terms of mean accuracy.

4. Discussion

Overall study of movie rating with machine learning techniques, Logistic Regression (82.53%) seems to be better compared with Naive Bayes algorithm (80.83%). There is a statistically insignificant difference in inaccuracy. It is more significant among decision and support vector machine algorithms ($p<0.05$ Independent Sample test) with a 95% confidence level.

The proposed work analyzed accuracy and precision of Logistic Regression and Naive Bayes algorithms for movie rating. Result shows evidence that there is statistical difference between Logistic Regression algorithms (97.2%) and Naive Bayes algorithm (93.2%) techniques. Logistic Regression algorithm accuracy appeared to be higher than Naive Bayes algorithm and compared accuracy with existing work [13]. The work describes how work has been done in the movie rating system and the difficulties faced while finding movies [14]. This method increases the performance by overcoming the problem that arises during the distribution of training data [15]. One research work described movie rating systems and showed their results using a particular algorithm with less accuracy [16]. The method was tested on a dataset of benchmark.

5. Conclusion

The strange approach of the movie rating system showed better results on the dataset. An early movie rating may be effectively performed using this system. Hence, the proposed system will serve as a significant tool for movie recommendation. The applied novel Logistic Regression algorithm and Naive Bayes algorithm arewell trained and tested. The testing is done using a cross validation method in which attributes were based on the movie dataset named ratings. The result obtained with a level of significance of0.003 proving that the Logistic Regression provides better result when compared to Naive Bayes algorithm.

References

[1] Sahu, Tirath Prasad, and SanjeevAhuja. 2016. "Sentiment Analysis of Movie Reviews: A Study on Feature Selection Amp; Classification Algorithms." In 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), 1–6.

[2] S.Kalaiarasi, P.Sriramya, "Optimization of Multiple Travelling Salesman Problem using a Novel Hybrid Ant Breeding Algorithm", International Journal of Advanced Science and Technology, 28(11), 2019

[3] Advani, Vaishali. 2020. "Movie Recommendation System in Machine Learning." April 7, 2020. https://www.mygreatlearning.com/blog/masterclass-on-movie-recommendation-system/Advani, Vaishali. 2020. "Movie Recommendation System in Machine Learning." April 7, 2020. https://www.mygreatlearning.com/blog/masterclass-on-movie-recommendation-system/

[4] Campos, Luis M. de, Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Miguel A. Rueda-Morales. 2010. "Combining Content-Based and Collaborative Recommendations: A Hybrid Approach Based on Bayesian Networks." International Journal of Approximate Reasoning. https://doi.org/10.1016/j.ijar.2010.04.001.

[5] Kumar, R., S. A. Edalatpanah, S. Jha, and R. Singh. 2019. "A Pythagorean Fuzzy Approach to the Transportation Problem." Complex & Intelligent Systems. https://doi.org/10.1007/s40747-019-0108-1

[6] S.Kalaiarasi, P.Sriramya, "Optimization of Multiple Travelling Salesman Problem using a Novel Hybrid Ant Breeding Algorithm", International Journal of Advanced Science and Technology, 28(11), 2019

[7] Srivastava, Suyash, Lokesh Sharma, Vijeta Sharma, Ajai Kumar, and Hemant Darbari. 2019. "Prediction of Diabetes Using Artificial Neural Network Approach." Engineering Vibration, Communication and Information Processing. https://doi.org/10.1007/978-981-13-1642-5_59

[8] Tripathi, Kshitij. 2019. "Diabetes Classification And Prediction Using Artificial Neural Network." International Journal Of Computer Engineering & Technology. https://doi.org/10.34218/ijcet.10.3.2019.018

[9] Fisher, Michael J., Allan J. Belzberg, Peter de Blank, Thomas De Raedt, FlorentElefteriou, Rosalie E. Ferner, Marco Giovannini, et al. 2018. "2016 Children's Tumor Foundation Conference on Neurofibromatosis Type 1, Neurofibromatosis Type 2, and Schwannomatosis." American Journal of Medical Genetics. Part A 176 (5): 1258–69.

[10] Campos, Luis M. de, Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Miguel A. Rueda-Morales. 2010. "Combining Content-Based and Collaborative Recommendations: A Hybrid Approach Based on Bayesian Networks." International Journal of Approximate Reasoning. https://doi.org/10.1016/j.ijar.2010.04.001

[11] Miyahara, Koji, and Michael J. Pazzani. 2000. "Collaborative Filtering with the Simple Bayesian Classifier." PRICAI 2000 Topics in Artificial Intelligence. https://doi.org/10.1007/3-540-44533-1_68.

[12] Han, Jiawei, Jian Pei, and MichelineKamber. 2011. Data Mining: Concepts and Techniques. Elsevier.

[13] Suleiman, Dima, and Ghazi Al-Naymat. 2017. "SMS Spam Detection Using H2O Framework." Procedia Computer Science. https://doi.org/10.1016/j.procs.2017.08.335

[14] Prasad, B. M. K., Karan Singh, Shyam S. Pandey, and Richard O'Kennedy. 2019. Communication and Computing Systems: Proceedings of the 2nd International Conference on Communication and Computing Systems (ICCCS 2018), December 1-2, 2018, Gurgaon, India. CRC Press.

[15] Dolgikh, Dmitry, and Ivan Jelínek. 2016. "Graph-Based Rating Prediction Using Eigenvector Centrality." Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. https://doi.org/10.5220/0006044902280233

[16] De Bra, Paul, Alfred Kobsa, and David Chin. 2010. User Modeling, Adaptation, and Personalization: 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010, Proceedings