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

doi:10.3233/APC220042

Unique Approach of Movie Rating Prediction Using Support Vector Machine Algorithm and K-Nearest Neighbors 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 predict movie rating using Support Vector Machine Algorithm and K-Nearest Neighbors Algorithm. An aggregate of 392 examples were gathered from film datasets and these datasets were taken fromkaggle dataset. Two experimental calculations were performed, one with K-Nearest Neighbors algorithm and another with Novel Support Vector Machine algorithm. Sample size of N=5 is taken for both algorithms. The computation processes were executed and verified for exactness. SPSS was used for predicting significance value of the dataset considering G-Power value as 80%. Novel Support Vector Machine calculation was applied and it had accomplished mean accuracy of 91.8% when compared with mean accuracy of 53% of K-Nearest Neighbors algorithm. The outcomes were obtained with a significance value of 0.03 ($p<0.05$). A unique approach model is affirmed to have higher exactness than K-Nearest Neighbors calculation.

Keywords: K-Nearest Neighbors Algorithm, Novel Support Vector Machine Algorithm, Collaborative, Machine Learning, Artificial Intelligence.

1. Introduction

In this busy world, entertainment plays a major role in refreshing each one of us, not only our mood but also our energy. By watching interesting movie, people get their mind relaxedby gettingout of routine work for a while. After that, they may once again rededicate themselves to the work.People listen to their favorite music or watch movies of their choiceto overcome their stress. To watch favorable movies online, movie recommendation systems can give the best suggestions. Searching for preferred movies require more time. A movie recommendation system that employs a hybrid approach by combining content-based filtering and collaborative filtering, using Novel Support Vector Machine as a classifier was applied in various Artificial Intelligence applications [1][2]. It requires a large amount of information about a user to make accurate recommendations.

Nowadays, with advancement in mobile technology, life has become easy and enjoyable [3]. Sentiment analysis paves a vital role in movie rating [4]. These systems have certain limitations, and they can operate using multiple inputs within and across platforms like news, books, and search queries [5]. The application of this research

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

work helps the person who looks for news, reviews, facts, trivia, and other details about your favorite films and TV series, on the Internet Movie Database. Moreover, with apps like BetaSeries, Cineast, and JustWatch, one can track down top movies of importance and it can even track upcoming movies.

There are 16 exploration articles published in IEEE Xplore and around 1700 articles found by Google researchers. In recent times, surveys of machine learning algorithms for movie ratings were explored mostly as they were predicted to have 80% of output accuracy. Kumar proposed a movie suggestion framework dependent on collective separating approaches [6]. However, it has less client criticism of customers on movies. This particular article is cited 20 times. This framework consolidates both collaborative and content-based techniques[7]. K-Nearest Neighboring (KNN) collaborative filtering algorithm is a combination of both the collaborative filtering algorithm and the KNN algorithm. KNN algorithm is used to select neighbors. Calculation steps are user likeness figuring, KNN nearest neighbor choice and anticipating score estimationare the main features of the algorithm [8]. This has certain limitations in providing the particulars, like number of fights, number of songs, etc. KNN is a simple idea that involves creating a distance metric between items in a dataset and then identifying the K things that are closest to it[9]. Based on observation of the above literature survey, accuracy and prediction of movie rating systems is less when using other algorithms. This proposed unique approach is used for better accuracy and precision. Thus, the aim of this research is to predict movie ratings with improved accuracy by using Novel Support Vector Machine algorithm over K-Nearest Neighbors algorithm.

2. Materials And Methods

The complete study was done in the Machine Learning lab at Saveetha School of Engineering. This examination comprises two example clusters i.e, one is the KNN algorithm and another is the Novel SVM algorithm, by using the python programming language. Each algorithm deals with 392 examples with a sample size N=5, a pretest power value of 80% taken for testing with an alpha value of 0.05. The Movie recommendation system proposes a framework to incorporate statistical data analysis for clustering techniques in data mining [10].

2.1 Novel Support Vector Machine Algorithm (NSVM)

Novel Support Vector Machine Algorithm is a technique of classification that separates the values of data by the creation of hyperplanes. Hyperplanes can be of different shapes based on the spread of data, but just points that help in distinguishing between classes are considered for classification. The novel Support Vector Machine algorithm imports python libraries and the Support Vector machine SVC library with kernel as linear. With the help of a function containing train labels and train features, it will predict output.

2.2 K-Nearest Neighbors algorithm

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric classification method that was developed by [11] and expanded later [12]. It is used for classification

and regression. In k-NN regression, the output is the property value for an object. This value is the average of values of k nearest neighbors.

The algorithm relies on distance for classification. If features represent different physical units, then normalizing training data can improve its accuracy dramatically. For effortlessness, this classifier is called a KNN Classifier. KNN classifier addresses example acknowledgment issues and most ideal decisions for tending to a portion of grouping-related errands.

The dataset for testing and training was collected and data preprocessing was completed initially. After that, split the dataset into a training set of 30% and testing set of 70%. Cross-validation needs to be done automatically, and the split function generates and implements machine learning classifiers. 30% trained dataset will train classifiers for the remaining part of 70% dataset will be trained using machine learning algorithms. After training, the classifier uses a testing dataset to check trained classifiers to urge anticipated accuracy from the classifier. The whole dataset is suitable for training; both algorithms and the accuracy of both models were tested with different sample sizes from 50 to 1000 as a step of 50.

2.3 Statistical Analysis

SPSS version 21 software dataset is prepared by using various samples and accuracy values obtained from the dataset are used for analysis. Here, data, and selection of movies are independent variables and accuracy, movie ratings are dependent variables. In SPSS, the dataset is prepared using 5 iterations from each of the algorithms. The group ID is given as 1, for KNN and Group ID is given as 2 for faster computation of SVM algorithm. Independent t-test analysis was carried out for this research study.

3. Result

In Table 1, it was observed that the dataset consists of 2 columns. Column 1 indicates the rating range from 0 to 5, and Column 2 indicates the location of the customer.

Table 1. Movie Ratings of Customers, the dataset column 1indicates the rating range, which is from 0 to 5, and Column 2 indicates the 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 N=5 samples of dataset for KNN and SVM algorithm with the highest accuracy of 58% and 96% in sample 5, using training data and testing data are respectively 30% and 70%.

Table 2. Performance comparison of algorithms with 5 iterations, N=5 sample size of the dataset for KNN and SVM algorithm with the highest accuracy of 58% and 96% in sample 5, using the training data and testing data of 70% and 30%, respectively.

S.NO	KNN Algorithm Accuracy %	SVM Algorithm Accuracy %
1	50	89
2	51	90
3	52	91
4	54	93
5	58	96

In **Table 3**, it was observed that t-test comparison, group Statistic analysis, represented KNN (mean accuracy 53.00%, SD=3.162) and SVM (mean accuracy 91.80%, SD=2.775).

Table 3. T-test comparison, Group Statistic analysis, representing KNN (mean accuracy 53.00%, SD = 3.162) and SVM (mean accuracy 91.80%, SD = 2.775)

Performance	Algorithm	N	Mean	SD	Error
Accuracy	KNN-Algorithm	5	53.00	3.162	1.414
	SVM-Algorithm	5	91.80	2.775	1.241

Table 4. Independent Sample Test for SVM and KNN (mean difference -38.800 and SD error difference1.881) provides statistical significance of 0.001 (2-tailed)

	Variance	F	Sig	t	df	Sig (2-tailed)	Mean difference	Std error diff	lower bound	upper bound
Accuracy	Equal variances assumed	0.062	0.03	-20.62	8	0.001	-38.80	1.881	-43.13	-34.46
	Equal variance not assumed			-20.62	7.86	0.001	-38.80	1.881	-43.13	-34.46

In **Table 4**, it was observed that independent samples test, for SVM and KNN (mean difference -38.800 and SD error difference 1.881 with significance 2-tailed 0.001 and respectively). F-is Fisher test, which is applied for testing of hypothesis, t-test is applied for comparing two groups with 95% confidence interval and df is degrees of freedom for n samples.From **Figure 1**, it was observed that this shows accurate ratings of movies. The following figure give us top 25 Movie ratings given to customers to visualize them 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 the rating levels given by customers.

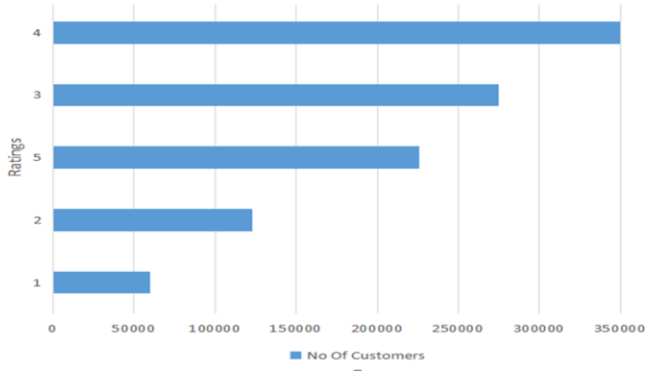

Figure 1. Top twenty five Movie Ratings given to the customer

From **Figure 2**, it was observed that it shows comparison of KNN-algorithms and SVM classifiers in terms of mean accuracy. The mean accuracy of SVM is better than KNN, and Standard Deviation of SVM is slightly better than KNN. In X Axis: SVM vs KNN Algorithms is shown and in the Y Axis: Mean accuracy of detection ± 1 SD is represented.

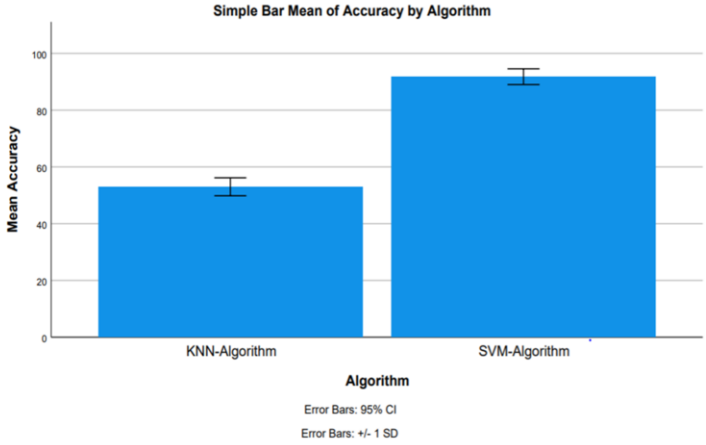

Figure 2. Bar graph of KNN-algorithm and SVM classifier

4. Discussion

Support Vector Machine algorithm (91.00%), has better accuracy compared with KNN algorithm (53%). Hence, it is inferred that the Support vector machine algorithm produces better accuracy than previous algorithms. There is a statistically insignificant difference in accuracy. It is more significant among Support Vector Machine and KNN algorithms ($p<0.05$, independent sample test) with a 95% confidence level. There is a significant difference in accuracy between two algorithms which is mentioned in group statistics and independent sample t-test.

The exactness of KNN calculation [13] and SVM calculation for 5 overlay cycles was referenced. In this examination, it is seen that KNN calculation demonstrated with better precision, appeared with box plot. The mean difference =

24.590 and the confidence interval is from -30.639 to -18.541. Subsequently, one can utilize it in a wide range of circumstances because unlabeled information can regularly be more accessible [14]. A large number were proposed for various use cases dependent on this execution; for instance, there are affiliation learning calculations that consider the requesting of things, their number, and related timestamps [15].Due to limitations such as relatively small size dataset, threshold, precision and recall reduce accuracy of movie rating [16]. Movie data used in this dataset is collected from various sources where the reliability of the data may vary. And the simple networks that are used may not provide a better outcome on larger data sets [17][18]. Moreover in KNN, mean error appears to be higher than in support vector machines. In future, the dataset with many attributes can be taken and thus helps classifier to work efficiently for improving the prediction accuracy.

5. Conclusion

The unique approach showed better results on the dataset. Novel SVM calculation was applied and it had accomplished mean accuracy of 91.8% that is higher than KNN. Outcomes were obtained with a degree of importance ($p<0.05$) pretest power-worth of 80% executed by using more than 392 sample values.

References

[1] Li, Yingxian, Junwu Xu, and Min Yang. 2021. “Collaborative Filtering Recommendation Algorithm Based on KNN and Xgboost Hybrid.” Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1748/3/032041.

[2] Sanyal, Ritabrata, Devroop Kar, and Ram Sarkar. 2021. “Carcinoma Type Classification from High-Resolution Breast Microscopy Images Using a Hybrid Ensemble of Deep Convolutional Features and Gradient Boosting Trees Classifiers.” IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM PP (April). https://doi.org/10.1109/TCBB.2021.3071022.

[3] Jannach, Dietmar, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. Recommender Systems: An Introduction. Cambridge University Press.

[4] Sahu, Tirath Prasad, and Sanjeev Ahuja. 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.

[5] Kapoor, Nimish, Saurav Vishal, and Krishnaveni K. S. 2020. “Movie Recommendation System Using NLP Tools.” In 2020 5th International Conference on Communication and Electronics Systems (ICCES), 883–88.

[6] 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

[7] De Raedt, Luc, KristianKersting, SriraamNatarajan, and David Poole. 2016. Statistical Relational Artificial Intelligence: Logic, Probability, and Computation. Morgan & Claypool Publishers.

[8] Ihsan, M. Arinal, M. Arinal Ihsan, Muhammad Zarlis, and Pahala Sirait. 2018. “Reduction Attributes on K-Nearest Neighbor Algorithm (KNN) Using Genetic Algorithm.” Proceedings of the 3rd International Conference of Computer, Environment, Agriculture, Social Science, Health Science, Engineering and Technology. https://doi.org/10.5220/0010043203710378

[9] Patra, Sukanya, and Boudhayan Ganguly. 2019. “Improvising Singular Value Decomposition by KNN for Use in Movie Recommender Systems.” Journal of Operations and Strategic Planning. https://doi.org/10.1177/2516600x19848956

[10] 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

[11] Fix, Evelyn, and J. L. Hodges. 1951. “Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties.” PsycEXTRA Dataset. https://doi.org/10.1037/e471672008-001

[12] Cover, T., J. Thomas, R. Marks Ii, S. Verdú, H. Poor, C. Looney, and Thomas Cover. 1997. "Information Theory." Electrical Engineering Handbook. https://doi.org/10.1201/9781420049763.ch73

[13] Shi, Yue, Martha Larson, and Alan Hanjalic. 2010. "Mining Mood-Specific Movie Similarity with Matrix Factorization for Context-Aware Recommendation." Proceedings of the Workshop on Context-Aware Movie Recommendation - CAMRa '10. https://doi.org/10.1145/1869652.1869658

[14] Hamza, Ali, Iftikhar Ahmad, and Muhammad Uneeb. 2021. "Fuzzy Logic and Lyapunov-Based Non-Linear Controllers for HCV Infection." IET Systems Biology 15 (2): 53–71.

[15] Sethi, Alakh. 2020. "Support Vector Regression In Machine Learning." March 27, 2020. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

[16] Prendin, Francesco, Simone Del Favero, Martina Vettoretti, Giovanni Sparacino, and Andrea Facchinetti. 2021. "Forecasting of Glucose Levels and Hypoglycemic Events: Head-to-Head Comparison of Linear and Nonlinear Data-Driven Algorithms Based on Continuous Glucose Monitoring Data Only." Sensors 21 (5). https://doi.org/10.3390/s21051647

[17] Frangi, Alejandro F., Jerry L. Prince, and Milan Sonka. 2019. Medical Image Analysis. Academic Press.

[18] Lu, Le, YefengZheng, Gustavo Carneiro, and Lin Yang. 2017. Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets. Springer.