Predictive Analytics on YouTube Videos Popularity using k-medoids Algorithm

Sangeetha T.G.R [1], Renuka Devi.D [2]

PG Scholar, Department of Computer Science, Stella Maris College, Chennai, India [1]

Assistant Professor, Department of Computer Science, Stella Maris College, Chennai, India [2]

Abstract: Nowadays YouTube programs attain more publicity and people are much addicted on these programs. In order to uncover the most popular program and the channel name we propose a novel method of predictive analytics. The accurate and sensible forecast about a program's popularity provides great value for people like content providers, advertisers, and broadcast TV operators etc. This information can be useful towards cost-effective investment plans. There are quite a lot of prediction models that are commonly used to predict program popularity. But these methods require abundant samples, extensive training and has poor prediction accuracy. An improved prediction approach is proposed and it uses the k-medoids algorithm for clustering the data into four trends and then it is given as input to gradient boosting decision tree and also in extreme gradient boosting algorithm. The prediction accuracy is also analyzed.

Keywords:. *YouTube, k-medoids algorithm, Predictive Analytics, Gradient Boosting*

*****

I. Introduction

As the eminence of new technologies like 3D technology increase, people get more obsessed to internet videos, so it attracted all the broadcast TV channels to bring out their programs in channels like YouTube etc. It is now becoming an emerging trend to telecast TV programs in internet to increase their popularity. According to the modern explorations the internet streaming of broadcast TV programs will continue to grow at a rapid pace. All the programs do not get equal response. Only a few programs can gain enormous user attention the remaining programs are left without anybody to watch them.

In this perspective, it is of great magnitude to forecast the popularity of these programs. Using the program popularity prediction results, the audience will save much time when trying to discover valuable programs among massive collections of video resources, which will improve user satisfaction. Based on program popularity data, a company will be able to maximize its marketing effect by choosing the programs with highest potential.

However, accurately predicting the popularity of broadcast TV programs, quality of the program and the interests of the audience is a difficult task. Last, there is a massive gap between the popularity evolutionary trends of different programs, which should be considered when designing the prediction model [5].An enhanced method to predict the program popularity among YouTube programs is proposed in this paper.

II. RELATED WORK

The program popularity prediction began with the news articles. It formed a new way for online content prediction. It introduced innovative methods to predict news comment volume and popularity of news articles such as those discussed by M. Tsagkias et al[1]. G. Szabo and B. A. Huberman proposed the method comparable to the previous one [2].It uses a long linear model to predict the data. R. Crane and D. Sornette [3] observed a Poisson method can depict the popularity gained by videos followed three popularity evolutionary trends. H. Pinto et al.[4] proposed a method that was applied on YouTube data to forecast popularity of web content, based on chronological information given by early popularity measures. Chengang zhu [5] proposed a new k-medoids algorithm along with random forest regression to predict popularity content of a broadcast TV channel and it predicts the name of the program which is well-liked. T. Chen and C. Guestrin [6] introduced a new algorithm named novel sparsity aware algorithm to handle the problems faced while using XGBoost. K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Zomaya[7] combined various algorithms and methods to eliminate the difficulties in producing a good price forecasting model. G. Gürsun, M. Crovella, and I. Matta[8] described a new method effective prediction of video sharing sites using ARMA model. K.Wang *et al[9]* introduced a storage planning scheme for wireless data. M. Ahmed, S. Spagna, F. Huici, and S. Niccolini[10] proposed a new model to predict content popularity efficiently than the previously used log-linear model. All the previous methods used focuses only on the general model to predict the popularity of a program and are ineffective to predict popularity among broadcast channel and in this work they proposed new method to predict popularity among broadcast TV channel programs.

III. METHODOLOGY

The main aspects of our work on popularity prediction are as follows:

First, we use *K*-medoids algorithm to cluster programs with similar popularity into 4 evolutionary trends. This approach provides more efficient outcomes than the previous methods that were used to delineate popularity evolutionary trends [5]. Secondly, we put up trend-specific prediction models using gradient boosting algorithm and in extreme gradient boosting algorithm and find out which one achieves higher overall predictive performance. The proposed model is shown in the figure 1.

The methodology includes three different algorithms. First k-medoids algorithm finds the new evolutionary trends. For program popularity there are different types of propagation trends. Each one has different level of features. We could get more efficient data if they are propagated. So in order to propagate those k-medoids is used as a replacement for of k-means clustering. It is said that if the cluster groups are more than four it will not provide accurate prediction model. So this paper uses four clusters and is used in Gradient boosting algorithm to find the predictions.

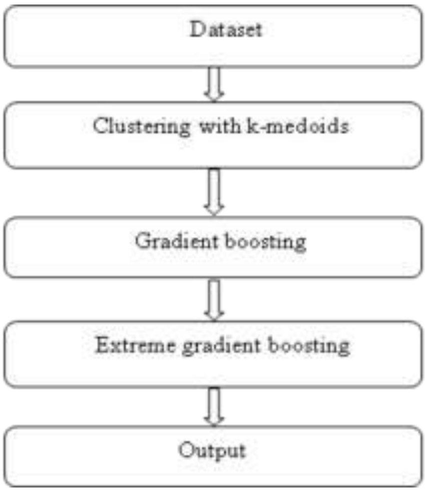

Fig-1.Proposed Model

Finally Extreme gradient boosting algorithm is used with the same input used in gradient boosting to predict the popularity of programs and check which one provides better predictions.

A.TREND DETECTION

This section describes the k-medoids clustering of program popularity into four trends [5]. In k-means clustering the center of the cluster represents the mean of the members in the cluster. In k-medoids the centre of the clusters are the medians of the cluster members. The k-medoids thus gives efficient results than k-means. The other steps of k-medoids algorithm are the similar to that of k-means. The clusters gained are more important and are used in different algorithms to predict the popularity.

B.TREND SPECIFIC PREDICTION

This section describes about the usage of Gradient boosting algorithm for providing trend specific prediction models. It produces more accurate values than the other prediction algorithms. The decision trees are perceptive to the data on which they are trained [5].The other algorithms have high structural similarities but in Gradient boosting the trees are unique. It is specified that stable results for estimating variable importance are achieved with a higher value [5].

C.CLASSIFICATION OF PROGRAM'S POPULARITY USING EXTREME GRADIENT BOOSTING

This section discusses about extreme gradient boosting algorithm. This algorithm is built based on the principles of gradient boosting. The difference between gradient boosting and extreme gradient boosting is that it produces a regularized model formulation to manage over-fitting, which produces a better performance. A solitary decision tree can have over fitting which is overcome by gradient boosting algorithm by combining hundreds of trees each containing some leaf nodes [5]. The extreme gradient boosting model gives better forecast presentation when compared with other models and it also has a great speed. It is ten times faster than other algorithms. The decision trees are built to predict new popularity trends. Thus it produces an efficient result on predicting popularity among YouTube videos.

.

IV. EXPERIMENTAL RESULTS

The data used in this paper is YouTube trending videos dataset. It includes several features like title of the program, channel name, views, likes, dislikes etc. The summary of dataset is given below in table1.

Table1: Summary of dataset

	Views	Likes	Dislikes
Min	1141	0	0
Median	357858	8774	276
Mean	1109764	41621	2073
Max	66637636	2542863	504340

The proposed work is implemented in R-studio for k-medoids clustering, gradient boosting and extreme gradient boosting algorithms in this study. First the k-medoids algorithm is used to split the data into 4 clusters. This is shown in the fig.2.

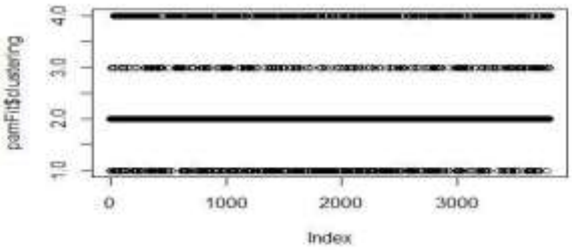

Fig-2.k-medoids clusters

The clusters gained from the above algorithm are used in Gradient boosting and the predictions are made according to these clusters. Then the clusters are used for Extreme gradient boosting model and predict the popularity for all the programs.

Table 2: Comparison of XGBoost and gradient boosting algorithm

Algorithm	MAE mean	RMSE mean	Rsquared mean
Gradient boosting	0.5143225	0.7384702	0.07186990
Extreme gradient boosting	0.5353799	0.7734377	0.03485341

Compared with gradient boosting algorithm the extreme gradient boosting algorithm gives better results. The comparison is shown in the table 2 and Fig 3.

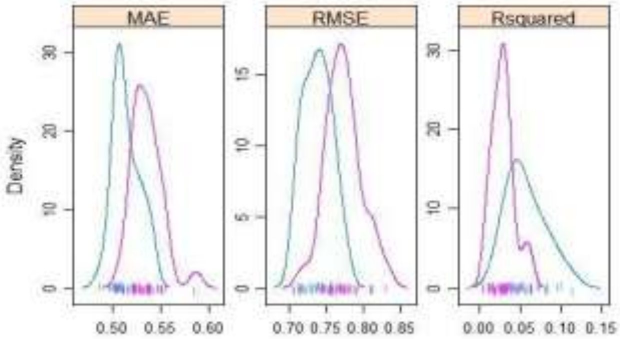

Fig-3.Comparison of gradient and extreme gradient algorithms.

V. CONCLUSION

In this paper we have predicted the popularity for programs using gradient and extreme gradient algorithms. We used *K*-Medoids algorithm to cluster programs into 4 trends, which has the capability to detain the program popularity. Furthermore, Gradient boosting is used to forecast the popularity. Then Extreme gradient boosting is used to predict the results. It gives more accurate prediction results than the generally used gradient boosting algorithm. The experimental results give gain in accuracy than the methods used previously to forecast program popularity among YouTube videos. It gives an unswerving prediction outcome much faster.

REFERENCES

[1] M. Tsagkias, W. Weerkamp, and M. de Rijke, ``News comments: Exploring, modeling, and online prediction,'' in *Advances in Information Retrieval*. Cham, Switzerland: Springer, 2010, pp. 191_203.

[2] G. Szabo and B. A. Huberman, ``Predicting the popularity of online content,'' *Commun. ACM*, vol. 53, no. 8, pp. 80_88, 2010.

[3] R. Crane and D. Sornette, ``Robust dynamic classes revealed by measuring the response function of a social system,'' *Proc. Nat. Acad. Sci. USA*,vol. 105, no. 41, pp. 15649_15653, 2008.

[4] H. Pinto, J. M. Almeida, and M. A. Gonç_alves, ``Using early view patterns to predict the popularity of YouTube videos,'' in *Proc. 6th ACM Int. Conf.Web Search Data Mining*, 2013, pp. 365_374.

[5] Chengang zhu, Guang cheng, (Senior Member, IEEE), and kun wang 2,3, (Senior Member, IEEE)" Big Data Analytics for Program Popularity Prediction in Broadcast TV Industries",IEEE,2017.

[6] T. Chen and C. Guestrin, ``XGBoost: A scalable tree boosting system, ''presented at the Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, 2016.

[7] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Zomaya, ``Robust big data analytics for electricity price forecasting in the smart grid,'' *IEEE Trans.Big Data*, to be published.

[8] G. Gürsun, M. Crovella, and I. Matta, ``Describing and forecasting video access patterns,'' in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 16_20.

[9] K.Wang *et al.*, ``Wireless big data computing in smart grid,'' *IEEEWireless Commun.*, vol. 24, no. 2, pp. 58_64, Apr. 2017.

[10] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, ``A peek into the future: Predicting the evolution of popularity in user generated content,'' presented at the Proc. 6th ACM Int. Conf. Web Search Data Mining, Rome, Italy,2013.