DOI: https://doi.org/10.24113/ijoscience.v5i7.165

Analysis of Feature Reduction Techniques for Online News Popularity Prediction

Shivangi Bhargava
P.G. Student
Department of Computer Science and Engineering
Maharana Pratap College of Technology
Gwalior, India
shivi8856@gmail.com

Dr. Shivnath Ghosh
Associate Professor
Department of Computer Science and Engineering
Maharana Pratap College of Technology
Gwalior, India
shivghosh.cs@gmail.com

***Abstract*: News popularity is the maximum growth of attention given for particular news article. The popularity of online news depends on various factors such as the number of social media, the number of visitor comments, the number of Likes, etc. It is therefore necessary to build an automatic decision support system to predict the popularity of the news as it will help in business intelligence too. The work presented in this study aims to find the best model to predict the popularity of online news using machine learning methods. In this work, the result analysis is performed by applying Co-relation algorithm, particle swarm optimization and principal component analysis. For performance evaluation support vector machine, naïve bayes, k-nearest neighbor and neural network classifiers are used to classify the popular and unpopular data. From the experimental results, it is observed that support vector machine and naïve bayes outperforms better with co-relation algorithm as well as k-NN and neural network outperforms better with particle swarm optimization.**

Keywords – Machine Learning, Classification, Popularity Prediction, Correlation Co-efficient, Accuracy.

1. INTRODUCTION

In the digital world, online news is primary source of information [1]. Various businesses are keen to know what will be the future demand of online visitors. Popularity prediction is useful in many applications like media advertising, estimation of movie revenue, traffic management, economic trends forecasting. Popularity prediction is hard to capture as it depends upon various factors like its topic, text, timing, article's position on the web page, language, similarity with world's event, same subject historical popularity, time from news publication, season of article popularity, relevance to the physical world popular events [2,3].

The large numbers of prediction methods for various types of web content are proposed in the research of latest years [4]. This research work is focused on prediction task.

Number of shares is one of the factors to determine popularity of news articles. In this work, it is intended to find the better model to predict the online popularity of news by using different machine learning techniques [5].

Prediction of the popularity is considered into 2 parts i.e. popularity prediction before publication of news and popularity publication after publication of news [6]. Mostly popularity prediction before news publication is considered for study. In the latest years, different types of prediction methods for different types of web information have been proposed [7]. This study focused on prediction of large visitors' attention to particular news articles, its reasons, evaluated methodologies, considered parameters and improved results.

In the current scenario, this paper have proposed methodologies which provide a way to predict whether an article will become popular or not. The objective of the paper is to maximize the rate of prediction of the article by minimizing and selecting the optimum features. Publishers can benefit by estimating the popularity of the news content and strategize accordingly by focusing on the features obtained as a result of this analysis. Further this paper is enhanced to make comparative analysis of multiclass (popular, Unpopular and Average) popularity prediction methods by considering parameters. To fulfill the objectives of this paper, dataset of 39,797 news articles are collected from UCI machine learning

repository which is a collection of Mashable's online news website [8]. Different machine learning algorithms are planned to implement on the dataset to evaluate and compare their performances.

1. Correlation Analysis

A bivariate analysis used for measuring the degree of association amongst two vectors say A and B is known as Correlation. In data mining, the value obtained after doing Correlation analysis varies between ±1. When this value is greater than 0, then a positive correlation exists and if this value is less than zero, then a negative correlation exists. If the value is 0, then the relationship between them is weak. For the proposed work that correlation value is selected whose value is positive one.

In this paper for feature selection Correlation Analysis is performed using Pearson, Spearman and Kendall coefficients which are explained in algorithm 1, algorithm 2, algorithm 3 and algorithm 4.

Algorithm 1: Pearson Correlation Analysis

Pearson correlation coefficient ρ is calculated by the formula as given below:

$$\rho = \frac{E[AD] - E[A]E[D]}{\sqrt{E[A^2] - (E[A])^2}\sqrt{E[D^2] - (E[D])^2}}$$

where:

A stands for the Attribute Vector

D stands for the Decision Vector

E[A] stands for the sum of the elements in A

Algorithm 2: Spearman Correlation Analysis

Spearman Correlation coefficient σ is calculated by the formula mentioned below:

$$\sigma = 1-(6\Sigma d_i^2)/n(n^2 - 1)$$

Where,

d_i stands for the difference between the ranks of variables P and Q

n stands for the sample size

Algorithm 3: Kendall Correlation Analysis

Kendall Correlation coefficient τ is calculated by the formula as given below:

$$\tau = (n_c - n_d)/(1/2n(n - 1))$$

Where,

d_i stands for the difference between the ranks of variables P and Q

n stands for the sample size

After doing Pearson Correlation by Algorithm 1, Spearman Correlation using Algorithm 2 and Kendall-rank Correlation by Algorithm 3, we get a list of attributes that satisfy the respective correlation criteria. After obtaining the three individual results which reduces the number of features using Algorithm 4 discussed below:

Algorithm 4: Attribute Selection after Correlation

```
procedure ATTRIBUTESELCTION(Dataset)
rows ← nrows(Dataset)
cols ← ncols(Dataset)
pearsonVector ← pearson(Dataset)
spearmanVector ← spearman(Dataset)
kendallVector ← kendall(Dataset)
for each i in 1:cols do
if pearsonVector[i]>0 AND spearmanVector[i]>0
AND kendallVector[i]>0 then
Selection ← true
else
Selection ← false
end if
end for
return dataset[,Selection]
end procedure
```

2. PARTICLE SWARM OPTIMIZATION

The basic process of the PSO algorithm is given by:

Step 1: (Initialization) Create random initial particles. For the PSO algorithm, the complete set of entities is represented by a string of length N.

Step 2: (Fitness) Measure the fitness of each particle in the population. This fitness value is used to optimize the result. In this algorithm global minimum to determine fitness function for the accuracy of detection.

Step 3: (Update) Calculates the speed of each particle.

Step 4: (Construction) For each particle, move to the next position.

Step 5: (Termination) Stop the algorithm if the termination criterion is satisfied; return to Step 2 otherwise.

PSO Algorithm

```
For every particle or jobs
Initialize jobs
end
Do
For each job
Calculate fitness value
If the fitness value is greater than the best fitness value
(pBest) in history
Then set current fitness value as the new pBest
End
Choose the job with the best fitness value of all the
particles as the gBest
For each job
Calculate particle velocity
Update job position in queue
```

End
While maximum iterations or minimum error criteria is not attained.

Calculation of fitness function

Each Particle's fitness function is calculated using pbest as well as gbest which is best position among entire group of particles.

In each generation velocity and position of each particle is updated using following equation

$$v_{new} = v_{old} + c1 * r1 * (pbest – present\_position) + c2 * r2 * (gbest – present\_position)$$

$$present\_position = present\ position + v_{old}$$

Where, v is the particle velocity
Present_position is the current particle (solution)
Pbest and gbest are defined as stated before.
r1 and r2 is a random number between (0,1).
c1, c2 are learning factors. usually c1 = c2 = 2.

3. PRINCIPAL COMPONENT ANALYSIS

PCA is a commonly used dimensionality reduction algorithm that could give us a less dimensional approximation to the original dataset, while preserving potential variability. It is a type of pattern recognition in the data. PCA is a powerful tool for data analysis. Once the data profiles and the compressed data have been found, the number of dimensions can be reduced without much information loss.

The PCA phases are as follows:

- Get data
- Subtract the average
- Calculates the covariance matrix
- Calculates the eigenvalues of the covariance matrix
- Select the components and create a feature vector
- Get the new record.

4. DATA CLASSIFICATION

The initial data set had 61 attributes. The data set is modified by adding a 62nd attribute which is Boolean, named 'Popular' and 'Unpopular'. This attribute decides the class label of the data set which is based on average of the number of shares which is explained in algorithm 5.

Algorithm 5: Deciding Class of Articles

procedure POPULARITY(shares)
sum ← 0
for each i in shares do
sum ← sum + i
end for
avg ← sum
length(shares)
for each i in shares do
if i ≥ avg then
popularity ← true
else
popularity ← false
end if
end for
end procedure

For binary classification algorithm 6 is performed and flow diagram is shown in figure 2.

Algorithm 6: Binary Classification

Input: D {dataset};
Output: Label {Popularity Label};
Step1: For each instance in D, do
Find feature vector (V)
Step 2: For each V do
Feature Reduction using Pearson, Spearman and Kendall coefficient
Step 3: Data classification using Hybrid Classifier split data in two halves and classify data using SVM and RF algorithm
Step 4: Determine the total class label
Find
True_positive (TP)
True_negative (TN)
False_positive (FP)
False_negative (FN)
Step 5: Find Performance Parameters
Step 6: Predict Article popularity Class as
if (class=1) Article= Popular State
else_if(class =0) Article=Unpopular State
end for

5. DATA SET DESCRIPTION

The dataset is taken from UCI machine learning repository [8]. This dataset is collected from popular news web site known as Mashable.com. It is preprocessed and donated on this UCI repository. Total 61 attributes are extracted from 39,797 news articles and these attributes describe different features of every article. These news articles are collected during 2 years of period, from January 7 2013 to January 7 2015.

6. RESULT ANALYSIS

A news article is popular or unpopular is predicted based on last column of dataset known as 'number of shares' of news article on social media. Threshold value is calculated on 'number of shares' attribute

using algorithm 5. The entire dataset is split into training and testing set.

The result analysis is performed by applying Co-relation algorithm, particle swarm optimization and principal component analysis. For performance evaluation support vector machine, naïve bayes, k-nearest neighbor and neural network classifiers are used to classify the popular and unpopular data.

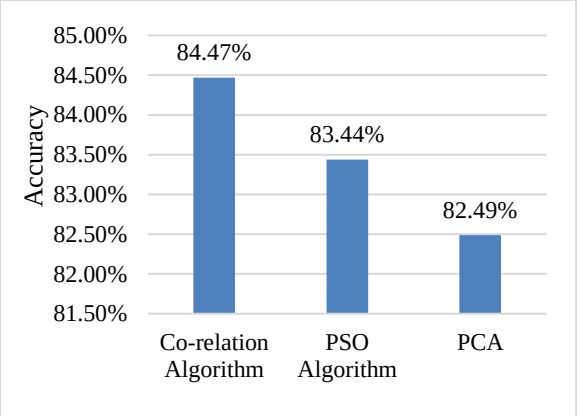

Figure 1: SVM Classifier Accuracy Evaluation

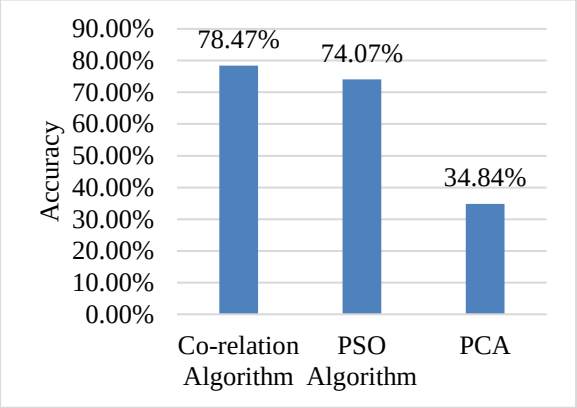

Figure 2: Naïve Bayes Classifier Accuracy Evaluation

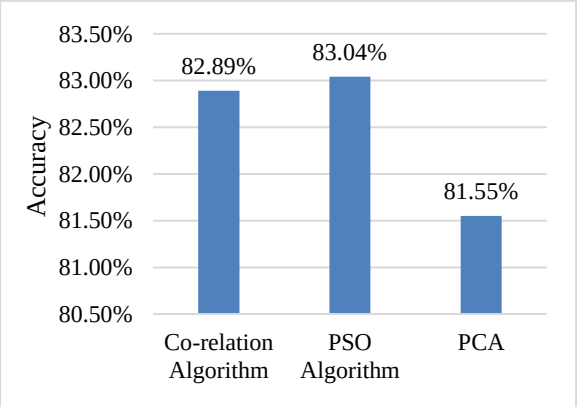

Figure 3: k-NN Classifier Accuracy Evaluation

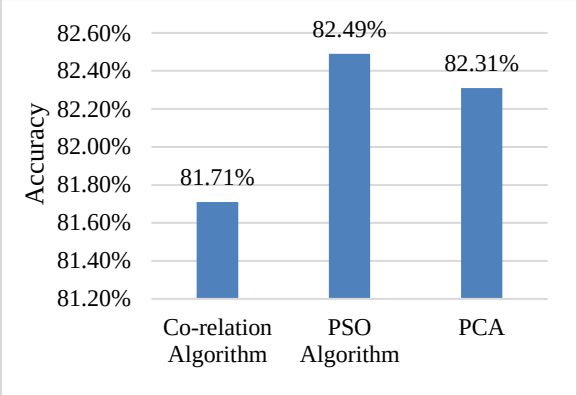

Figure 4: NN Classifier Accuracy Evaluation

7. CONCLUSION

News popularity is the maximum growth of attention given for particular news article. Online news popularity depends upon various factors such as number of shares on social media, number of comments by visitors, number of likes etc. So it is necessary to build an automated decision support system to predict the popularity of news as it will help in business intelligence too. The work presented in this research intends to find the best model to predict the popularity of online news by using machine learning methods.

After applying feature reduction, out of 61 attributes 30 attributes are reduced. So, reduced features are taken into consideration for prediction of popularity. In this work, performance evaluation metrics such as accuracy values are increased and given better performance of classification that is compared with existing research's implemented methodology. The machine learning methods like Support vector machine, Naïve Bayes, KNN and neural network is analyzed. From the experimental results, it is observed that support vector machine and naïve bayes outperforms better with co-relation algorithm as well as k-NN and neural network outperforms better with particle swarm optimization. So, in future work by analyzing this result, the work is proceeded for multiclassification.

REFERENCES

[1] Ilias N. Lymperopoulos, "Predicting the popularity growth of online content", Elsevier, Vol. 369, pp. 585-613, 10 November 2016.

[2] Kelwin Fernandes, Pedro Vinagre, Paulo Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News", Springer, EPIA 2015, pp. 535-546, 2015.

[3] He Ren, Quan Yang, "Predicting and Evaluating the Popularity of Online News", Standford University Machine Learning Report.

[4] Bandari Roja, Sitaram Asur, and Bernardo A. Huberman. "The pulse of news in social media: Forecasting popularity." arXiv preprint arXiv:1202.0332, 2012.

[5] Ioannis Arapakis, B. Barla Cambazoglu, and Mounia Lalmas, "On the Feasibility of Predicting News Popularity at Cold Start", Springer, pp. 290-299, 2014.

[6] R. Shreyas, D.M Akshata, B.S Mahanand, B. Shagun, C.M Abhishek, "Predicting Popularity of Online Articles using

Random Forest Regression", International Conference on Cognitive Computing and Information Processing, IEEE, 2016

[7] Swati Choudhary, Angkirat Singh Sandhu and Tribikram Pradhan, "Genetic Algorithm Based Correlation Enhanced Prediction of Online News Popularity" Computational Intelligence in Data Mining, Advances in Intelligent Systems and Computing, Springer, 2017, pp.133-144.

[8] UCI Machine Learning Database, https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity, May 2015.