International Journal of Information Science and Management
Vol. 16, No. 1, 2018, 71-87

Developing a New Hybrid Intelligent Approach for Prediction Online News Popularity

Jalal Rezaeenour
Associate Prof., Department of Industrial Engineering, University of Qom, Qom, Iran.
Corresponding Author,
j.rezaee@qom.ac.ir

Mansoureh Yari Eili
PhD candidate, Department of Computer Engineering and IT, university of Qom, Iran
m.yari@stu.qom.ac.ir

Esmaeil Hadavandi
Assistant Prof., Department of Industrial Engineering, Birjand University of Technology, Birjand, Iran.
es.hadavandi@birjandut.ac.ir

Mohammad Hossein Roozbahani
PhD candidate, department of Mechanical engineering, Tarbiat Modarres University, Tehran, Iran,
roozbahani.m.h@gmail.com

Abstract

This study aims to predict the amount of attention news articles ultimately receive using data mining technology. As well known, useful knowledge in Online Social Networking Services (Such as Digg, Twitter, Facebook and YouTube) is often hidden in large amounts of web data. Generally, due to dimensionality, irrelevant attributes will deteriorate the performance of the learning algorithms which increases training and testing times. In this paper, to reduce this impact in predicting the popularity of online news, a new feature selection algorithm is proposed based on Mutual Information. Cellucci-Mutual Information-based Feature Selection (MIFS) is firstly employed to select the most informative variables which affect the popularity of a news article. Then the selected features are used to train an Extreme Learning Machine (ELM) neural network. Experimental tests using practical datasets from the UCI repository were implemented to validate the performance of the proposed model. The analyses demonstrate that the proposed method can extract the most important features of online news data and can accurately predict future popularity. The prediction accuracy of ELM can improve dramatically using C_MIFS. With error rates RMSE=0.16 and MAPE=0.23. Hence, the new data mining model can provide practical application for online content popularity forecasting for digital media websites.

Keywords: Online Content Popularity Forecasting, Mutual Information based Feature Selection (MIFS), Extreme Learning Machine, Neural Networks, Prediction Method, Feature Selection

Introduction

In recent years, the advances in global availability of the internet and Online Social Networking (OSN) sites, such as YouTube, Twitter, and Facebook has caused interest in online news to surge. Due to the continuous flow of information and news around the globe, news articles have become extremely dynamic with low costs as well as short lifespans and sizes (Bandari, Asur, and Huberman, 2012; Tatar, Antoniadis, Amorim, and Fdida, 2014).

A key issue in producing online content is to predict online news popularity, which plays an important role in online advertisement and content distribution. For instance, being able to

predict which articles are likely to become popular can give competitive advantages to news sites and news aggregators. Furthermore, online readers can filter through huge amounts of information easily. Thus, this ability is also valuable for social scientists interested in understanding reader behavior (Tatar, Antoniadis, Amorim, and Fdida, 2014; Hensinger, Flaounas, and Cristianini, 2013)

The news article's position on the web page, timing, especially its topic and content have been regarded as important features in its popularity (public attention) (Bandari, Asur, and Huberman, 2012; Hensinger, Flaounas, and Cristianini, 2013). However, it is a significant challenge to accurately predict online popularity of news articles. However massive amounts of data which are continually generated on web sites make this prediction difficult.

It has been estimated that the amount of data stored in the databases around the world grows every twenty thousand at a rate of 100% (Vigneswaran, Joseph, and Rajamanickam, 2014). YouTube has reported that 24 hours of video content is uploaded to its servers every minute 5.

As a result, in such application, data are accumulating at increasing rates. The data are characterized by dynamic change, high volume, high dimensionality and potentially infinite length. These sets of data contain not only useful variables but also redundant, uninformative and even noisy information. Therefore, extracting useful information from these data can be very challenging.

So the study of content popularity prediction is significant. Experiments have shown that irrelevant attributes will deteriorate the performance of the learning algorithms for the curse of dimensionality, which increases training and testing times (Hoque, Bhattacharyya, and Kalita ,2014). Using data mining tools, it is possible to find potential patterns of dynamically large amounts of data available online and enhance the popularity forecasting (Frenay, Doquire, and Verleysen, 2013).

Feature selection is an important technique in machine learning and data mining since usually there are many candidate attributes collected to represent regression and classification problems. The objective is to find a few essential features from the original dataset (Hoque , Ahmed, Bhattacharyya, and Kalita, 2016; Foithong, Pinngern, and Attachoo, 2012). The selected inputs with minimum redundancy have maximum relevance with the output variables (Wang, Li, and Li, 2015).

In fact there are many potential benefits in feature selection, such as faster and more cost-effective input variables with more generalization capability, avoiding the curse of dimensionality and storage requirement, reducing computational cost and guaranteeing high accuracy and efficiency, data minimization and speed up learning process for a model (Huang, and Chow 2005; Long, Li, Fan, Xu, and Liang, 2013; Hacine-Gharbi, Ravier, Harba, and Mohamadi, 2012).

A common pertinence measure within feature selection methods is Shanon's Mutual Information (MI), I(c,x) between the class label C and the feature vector x (Shanon, 1948). Unlike correlation that only points out linear dependency, MI measures any type of statistical dependency between two variables.

MI can be considered as advanced statistics to rank salient features. When applying MI in feature selection, it plays a key role in measuring relevance and redundancy among features. Battiti was the first to apply MI to inputs and outputs for a classification problem (Battiti, 1994). MI-based Feature Selection (MIFS) selects the features that minimize the information of the class.

Corrected by subtracting a quantity proportion from the average MI with the previously selected features (Wang, Li, and Li, 2015).

The aim of this paper is to develop a new methodology for online content forecasting. To this end, a new MI-based feature selection scheme is introduced to quantify the effect of different factors that contribute to making an article news worthy. Moreover we use a learning algorithm for SLFN called Extreme Learning Machine (ELM) for prediction probability.

The remainder of this paper is organized as follows. In Section 2, related works are briefly reviewed. Section3 describes the background knowledge of MI and the Feature Selection based Mutual Information (MIFS) approach. In Section 4, the proposed C-MIFS algorithm is introduced. Details of the dataset, candidate features and experimental results are given in Section 5. Finally, concluding remarks are presented in Section 6.

Previous works

Predicting the popularity of news articles is a complex and difficult task with different prediction methods and strategies being proposed in several recent studies (Lee, Moon, and Salamatian, 2010; Lerman, and Ghosh, 2010; Szab, and Huberman, 2010). This section covers other works in the field.

In the past decades, a great number of studies have been performed to make use of an item's popularity to predict its future success(Tatar, Leguay, Antoniadis, Limbourg, Amorim, and et.al 2011; Kim, Kim, and Cho, 2011). These includes media advertising (Figueiredo, Benevenuto, and Almeida, 2011; Lakkaraju, and Ajmera, 2011), election prediction(Balasubramanyan, Routledge, and Smith, 2010 ; Tumasjan, Sandner, and Welpe, 2010; Williams, and Gulati, 2008), understanding user behavior (Crane, and Sornette, 2008; Kwak, Lee, Park, and Moon, 2010; Yang, and Leskovec, 2011), movie revenue estimation (Ahmed, Spagna, Huici, and Niccolini, 2013), and popularity of online content (Tsagkias, Weerkamp, and Rijke, 2009). The last application, however, has gained most of the research focus; predicting such popularity is valuable for authors, content providers, advertisers and even activists/politicians (e.g., to understand or influence public opinion). Popularity of YouTube videos is studied by (Lerman, and Ghosh, 2010). (Szabo, and Huberman, 2010) Looks at Digg stories and the social components that can contribute to their popularity. New article headlines are studied in (McCreadie, Macdonald, and Ounis, 2010) in terms of using blog information.

There are different ways of expressing the notion of popularity. For example, the classical way of defining it is by click-through rate. However, this information is rarely available to external observers and, when available, it is difficult to estimate the actual number of times that a page was requested by distinct users or by web crawlers and search engines.

Nevertheless, as reading news has become a social experience, there are other metrics that capture readers' interest. Popularity is often measured by metrics based on user participation interactions such as the number of shares, likes and comments, votes, click trough rates in the Web and social networks (Fernandes, Vinagre, and Cortez, 2015).

According to (Tatar, Amorim, Fdida, and Antoniadis, 2014) there are two main popularity prediction approaches: those that use features only known after publication and those that do not use such features. The first approach is more common (Ahmed, Stella Spagna, Huici, and Niccolini, 2013; Bandari, Asur, and Huberman, 2012; Kaltenbrunner, Gomez, and Lopez, 2007; Tatar, Antoniadis, Amorim, and Fdida, 2014). Since the prediction task is easier, higher prediction accuracies are achieved. In the second approach, lower prediction performance might be expected.

(Bandari, Asur, and Huberman, 2012) using the second approach, focused on four types of features (news source, category of the article, subjectivity language used and names mentioned in the article) to predict the number of tweets that mention an article. Four classification methods were applied on dataset from Feedzilla to predict three popularity classes (1 to 20 tweets, 20 to 100 tweets, more than 100; articles with no tweets were discarded). The results ranged in to about 84% accuracy. Hensinger (Hensinger, Flaounas, and Cristianini, 2013) using SVM obtained better results about 86% accuracy in data related ten English news outlets related with one year.

Szabo (2010) presented a linear regression to predict the long term popularity of an online content from early measurement of user access pattern. They observed a linear correlation between the logarithmically transformed long-time popularity of a content with the logarithm of its early measured popularity. Because the logarithmic transform was applied to popularity, the prediction errors behave as a multiplicative coefficient of the -term popularity. This results in large prediction errors. The weakness of linear regression is also confirmed by (Wu, Timmers, Vleeschauwer, and Leekwijck, 2010) where a reservoir computing-based prediction of the logarithm of the long-term popularity is proposed.

Most studies have focused on predicting the exact amount of attention that online content will generate in the future (Lee, Moon, and Salamatian, 2010; Tsagkias, Weerkamp, and Rijke. 2010). This information can indeed prove valuable in online advertising, where new revenue models could be designed to charge advertisers for the (future) amount of attention that a content will generate

Overall, the experiments confirm three points: (1) as efficient business intelligence tools, data mining and machine learning methods provide alternative tools to dynamically process huge amounts of data available online. (2) Comparable or even better performance can be achieved with reduced dimensionality of dataset. (3) The publication time of news has an important effect on reader attention, which has not studied before.

In this work, we aim to detect and identify the most important parameters which increase popularity patterns of online content. We show that the parameter of time has a direct impact on user behavioral patterns. Contrary to previous works, we build a model based on novel data mining techniques to predict future popularity of content.

Basic theory and algorithms

Theory of mutual information

This section gives a brief overview of basic definitions about mutual information and entropy. MI theory introduced by Shannon in 1948, and is an intuitive tool to measure the uncertainty of random variables and the amount information they share. A larger MI means that two variables share a larger extent of information.

Given two time series$S = \{S(t_1), S(t_2), \dots, s(t_N)\}$, and Q $= \{Q(t_1), Q(t_2), \dots, Q(t_N)\}$, their mutual information, I(S,Q), is the average number of bits of S that can be predicted by measuring Q. and can be expressed as:

$$I(S;Q) = H(S) + H(Q) - H(S,Q) \quad (1)$$

Where H (Q) and H(S) are the entropy of Q and S, respectively, and H(S, Q) is the mutual entropy between S and Q. Normally, the expressions of H(S), H (Q) and H(S, Q) are shown below.

$$H(Q) = -\sum P(q_i) log_2 P(q_i) \quad (2)$$

$$H(s) = -\sum P(s_i) log_2 P(s_i) \quad (3)$$

$$H(Q,S) = H(S,Q) = -\sum P(s_i, q_i) log_2 P(s_i, q_i) \quad (4)$$

Where P (si) is the probability distribution of section si. Eq. (1) changes to Eq. (2) as well.

$$I(Q,S) = \sum_i \sum_j P_{s,q}(s_i, q_i) log_2 \frac{P_{sq}(s_i, q_i)}{P_s(s_i) P_q(q_i)} \quad (5)$$

The relationship between H(S), H (Q), H(S, Q), H (S|Q), H (Q|S), and I(S; Q) can be expressed in a Venn diagram (Fig. 1).

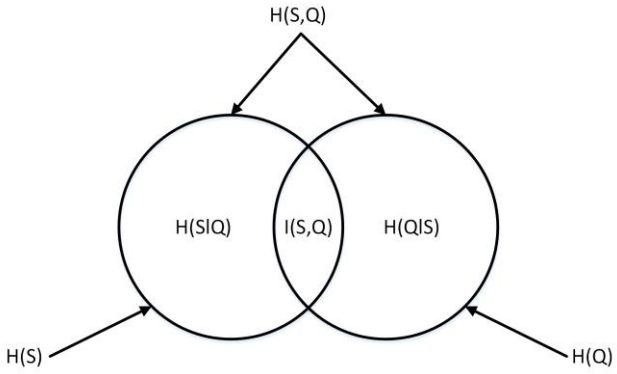

Figure 1. The relationship between entropy and mutual information

MI-based feature selection approach

Incidentally, in the study of feature selection, researchers mainly concentrate on measurement criteria and searching strategies. Among the different criteria, the information metric seems to be more comprehensively studied (Yan, Yuan, Yan, and Yang, 2011).

As mentioned before, the MI between two series, S and Q, is the average number of bits of S that can be predicted by measuring Q. In the analysis of observational data, MI is calculated in three contexts: (1) identification of nonlinear correlation; (2) determination of an optimal sampling interval, particularly when embedding time series data; and (3) in the investigation of causal relationships with directed mutual information.

The MIFS algorithm was proposed by Battiti in 1994 (Battiti, 1994). and focuses mainly on maximizing the MI between the candidate variable and the response variable and minimizing redundancy between candidate variable and selected variables. The process is shown in fig 2.

C-MIFS

The MI used in this paper is proposed by Cellucci (Cellucci, and Albano, 2005). Cellucci's algorithm posits that series S and Q are statistically independent. Base on this, the MI between two series can be obtained as below:

$$I(X,Y) = \sum_{i=1}^{N_E}\sum_{j=1}^{N_E} P_{XY}(i,j) ln\{N_E^2 P_{XY}(i,j)\} \quad (6)$$

Normally, N_E is taken as the maximal integer, which satisfies Eq :

$$N_E \leq \left(\frac{N}{5}\right)^{1/2} \quad (7)$$

Where N is the length of the series.

So, a new feature selection based Cellucci algorithm, called C-MIFS, is proposed. The procedure is described in the following:

MIFS based Cellucci Algorithm

Input: Set F of n features
Output: Set S of k features
For feature $f \in F$
Calculate $I(c;f)$
End
Find first feature f that maximizes $I(c;f)$;
Set $F \leftarrow F\backslash\{f\}$
Set $S \leftarrow \{f\}$
While $|S| < k$ do
Choose feature f as the one that maximizes:

$$I(X,Y) = \sum_{i=1}^{N_E}\sum_{j=1}^{N_E} P_{XY}(i,j) ln\{N_E^2 P_{XY}(i,j)\}$$

End

Proposed C-MIFS algorithm

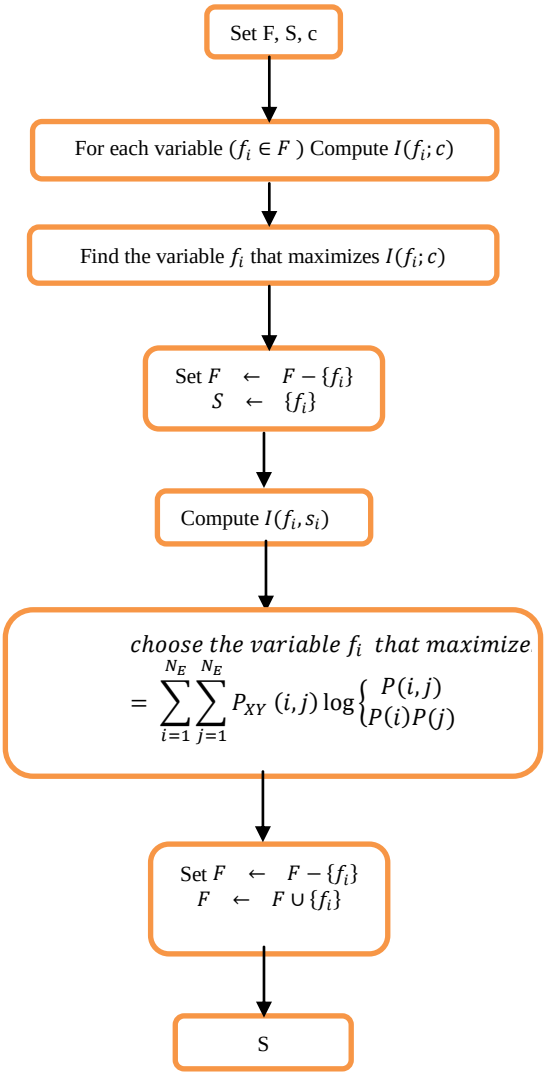

Fig2. The flowchart of C_MIFS algorithm

Extreme learning machine

This section briefly reviews ELM, originally proposed by Guangbin Huang (Huang, Zhu, and Siew, 2006).The main concept behind ELM lies in the random initialization of the SLFN weights and biases. Therefore, the input weights and biases do not need to be adjusted, which makes it possible to explicitly calculate the hidden layer output matrix and hence the output weights. Fig1. Shows an ELM architecture. Consider a set of M distinct samples $(x_i, y_i)\ with\ x_i \in \mathbb{R}^{d_1}$ and$y_i \in \mathbb{R}^{d_2}$; then, a SLFN with N hidden neurons is modeled as the following sum:

$$\sum_{i=1}^{N} \beta_i f\left(w_i^T X_j + b_i\right), \qquad 1 \le j \le M \tag{8}$$

With f being the activation function,w_i the input weights, b_i the biases and β_i the output weights. ELM is constructed in a way that it perfectly approximates the given output data:

$$\sum_{i=1}^{N} \beta_i f\left(w_i^T X_j + b_i\right) = y_j, \quad 1 \le j \le M \tag{9}$$

Which writes compactly as HB = Y, with

$$H = \begin{pmatrix} f(w_1 X_1 + b_1) & \dots & f(w_N X_1 + b_N) \\ \vdots & \ddots & \vdots \\ f(w_1 X_M + b_1) & \dots & f(w_N X_M + b_N) \end{pmatrix}_{N*M} \tag{10}$$

$$\beta = \begin{bmatrix} \beta_{11} & \dots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{M1} & \dots & \beta_{mM} \end{bmatrix}_{M*m} = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_M^T \end{bmatrix}_{M*m} \tag{11}$$

$$T = \begin{bmatrix} t_{11} & \dots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{M1} & \dots & t_{mM} \end{bmatrix}_{M*m} = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N*m} \tag{12}$$

H is called the hidden layer output matrix of ELM (Rezaeenour, Yari Eili, Roozbahani, Ebrahimi, 2016). The objective function for training the ELM is:

$$min\|T^{\cdot} - T\| = min\|H\beta - T\| \tag{13}$$

The solution provided in Eq. 13 sets the values of weights and biasesw_i, b_i which, given the hidden weights and biases previously established, minimize the mean square error (Lahoz, Lacruz, Pedro, and Mateo, 2013).

$$MSE(t_i, w_i) = \frac{1}{p} \sum_{i=1}^{n} \sum_{i=1}^{n} (T_i - f(w_i)^2) \tag{14}$$

The output functions of the hidden nodes may not be unique. Different output functions may be used in different hidden neurons. Particularly, in real applications *hi(*x*)* can be:

$$h_i(x) = G(a_i, b_i, x), a_i \in R_d, b_i \in R \tag{15}$$

Where *G (*a, *b*, x*)* (with hidden node parameters *(*a, *b)*) is a nonlinear piecewise continuous function satisfying ELM universal approximation capability theorems (Huang, Zhu, and Siew, 2006). Table1 represents the commonly used mapping function in ELM neural networks.

However, ELM tends to have problems when irrelevant or correlated variables are present. For this reason, OP-ELM methodology proposes a pruning of the irrelevant variables, via pruning of the related neurons of the SLFN built by the ELM (Cao, Lin, and Huang, G.B., 2010).

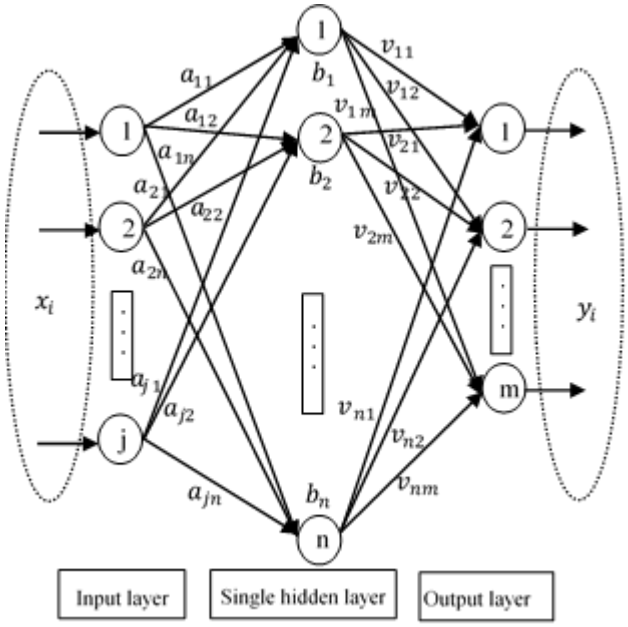

Figure 3. The structure of ELM

Table1
Commonly used mapping function in ELM

Sigmoid Function	$G(a,b,x)=\frac{1}{1+exp\left(-(a.x+b)\right)}$
Hyperbolic tangent Function	$G(a,b,x)=\frac{1-exp\left(-(a.x+b)\right)}{1+exp\left(-(a.x+b)\right)}$
Gussian Function	$G(a,b,x)=\exp(-b\|\|x-a\|\|)$
Multiquadric Function	$G(a,b,x)=\ (\|\|x-a\|\|+b^2)^{1/2}$
Hard Limit Function	$G(a,b,x)=\begin{cases}1, & if\ a.x+b\le 0\\ 0, & Otherwise\end{cases}$
Cosin Function/ Fourier basis	$G(a,b,x)=\cos(a.x+b)$

1- Experimental results

Dataset description

A real-world dataset was collected from the University of California at Irvine (UCI) Machine Learning Repository (http://archive.ics.uci.edu/ml/) by Kelwin Fernandes (Fernandes, Vinagre, and Cortez, 2015), which is used in literature for predicting popularity by machine learning algorithms. The term "popularity" for the news article refers to the number of times a new URL is posted and shared on Twitter (as an implicit indicator of the interest generated by a news article).

This dataset contained all the published articles during the period of January 7 2013 to January 7 2015 from Mashable, which is one of the largest news websites, with 39000 articles and a total of 47 features. The general information and details of the dataset are shown in Table 1. The interested reader is referred to (Fernandes, Vinagre, and Cortez, 2015)for specific technical details and more detailed information.

The attribute types were classified into: number – integer valued; ratio – within [0, 1]; bool – $\in \{0, 1\}$; and nominal. Column Type shows in brackets (#) the number of variables related with the attribute (Fernandes, Vinagre, and Cortez, 2015), the name, number of instances, number of features, and type of features are given in Table2.

Table2
News article features used in this study

Feature Type	Words	Number of words in the title number Number of words in the article number Average word length number Rate of non-stop words Rate of unique words Rate of unique non-stop words
	Links	Number of links number Number of Mashable article links Minimum, average and maximum of shares of Mashable links
	Digital Media	Number of images Number of videos
	Time	Day of the week Published on a weekend?
	Keywords	Number of keywords Worst keyword (min./avg./max. shares) Average keyword (min./avg./max. shares) Best keyword (min./avg./max. shares) Article category (Mashable data channel)
	Natural Language Processing	Closeness to top 5 LDA topics Title subjectivity Article text subjectivity score and its absolute difference to 0.5 Title sentiment polarity Rate of positive and negative words Pos. words rate among non-neutral words Neg. words rate among non-neutral words Polarity of positive words (min./avg./max.) Polarity of negative words (min./avg./max.) Article text polarity score and its absolute difference to 0.5
	Target	Number of article Mashable shares

Table3
Description of the dataset

Dataset	Service Address	Start-End Duration	Number of Articles
Online news popularity dataset	www.Mashable.com	January 7 2013-January 7 2015 (709 days)	39646

Performance evaluation

This section evaluates the proposed model in terms of prediction performance and the number of features in order to determine whether C-MIFS is adequate for a big dimensional dataset.

Following feature selection with the proposed C-MIFS, prediction performance with proposed input variable selection algorithm evaluating numerical computations. Our findings are shown in several tables and plots.

All simulations were conducted in the Mat lab R2010b environment running on a PC with a 2.5 GHz Core™ i5 CPU and 6 GB of RAM. Fig. 2 shows the MI between the target feature (number of shares) and the other features calculated by the proposed C-MIFS method. This plot is the MI diagram I(F,C) obtained for each feature in the database .

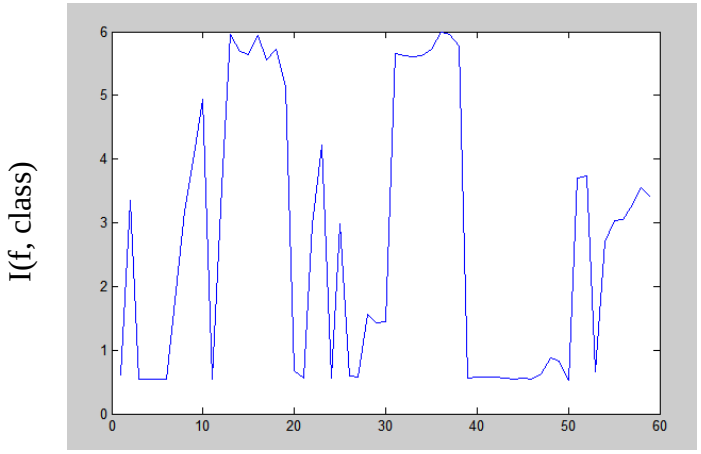

Figure 4. MI diagram for the features in dataset calculated by C-MIFS

Each feature gained a value between zero and six, revealing the importance of that feature with regards to a target feature. Features with weights higher than three were selected. A total of 20 informative features which influence article popularity were identified. Table 4 shows the description of the most effective features selected by the CMIFS algorithm.

Table 4
Description of the most effective features selected by the CMIFS algorithm.

Features	I(C,F)
Number of words in the title, Number of keywords in the metadata, Number of links, Min. polarity of positive words Max. polarity of positive words, Max. polarity of negative words title_ subjectivity: Title subjectivity Title polarity Absolute subjectivity level Absolute polarity level	[3,4]
Number of images, Number of videos, Min. polarity of negative words	[4,5]
Is data channel 'Entertainment'? Is data channel 'Lifestyle'? Is data channel 'Business'? Is data channel 'Social Media'? Is data channel 'Tech'? Is data channel 'World'? Worst keyword (min. shares) Worst keyword (max. shares) Which day the article has published? (On a Monday, Tuesday… or on the weekend?	[5,6]

As seen, the most important features are news in lifestyle and social media sections and publishing the news on the weekends (Saturdays and Sunday) which obtained maximum rate of MI (approximately 6). Therefore, article popularity is mainly associated these feature

Then these data with 20 features and 39000 samples are tested on ELM neural network, for prediction. The activation function of ELM is sigmoid function: $g(x) = 1/(1 + exp(-\alpha x))$, where α is 0.5. The dataset was divided into two parts for training and test purposes. Two thirds of instances are taken for the training set, and the remaining third for the test set.

In the experiments, dataset with different number of selected features extracted by different algorithms were tested on ELM neural network. In Fig.5 predictions of the dataset are displayed. As evident, once trained, the model is quite capable of estimating popularity with satisfactory accuracy.

As a measure of prediction performance of each feature extraction method, we adopt two measures: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), which are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_T - x_P)^2} \quad (16)$$

$$MAPE = \sum_{i=1}^{n}\frac{1}{N}\left|\frac{x_T - x_P}{x_T}\right| \quad (17)$$

where N is the number of testing data, x_T is the actual output value and x_P is the estimated output value, for $i = 1,2, \ldots, N$ respectively.

Results of the experiments, training and testing instance, and number of hidden nodes of the ELM are given in Table5. The computational time of GELM method for the dataset is about 15min on a corei5 CPU.

For all algorithms, the best RMSE is shown in bold face. It is clearly proved that the C-MIFS produces efficient and suitable results compared to popular feature selection methods. The proposed method produces suitable results with respect to accuracy as well as RMSE error.

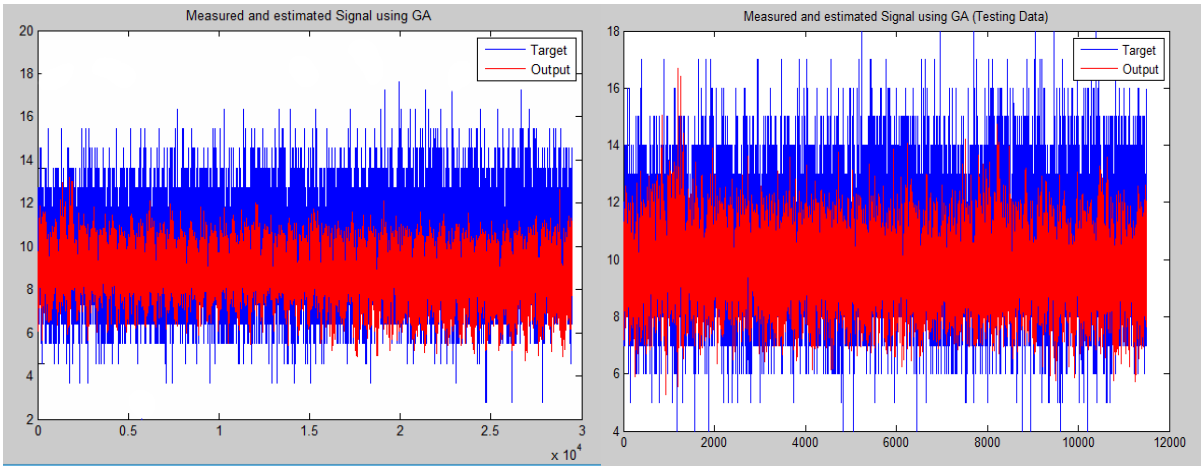

Fig5. Forecasting values and detected values of ELM model

Table5

Experimental results of ELM

	Number of training instance	Number of testing instance	Number of hidden nodes	RMSE	MAPE	Computational Time(s)
ELM	28000	11000	3	0.3848	0.23	1.07e+3

In the experiments, datasets with different number of the subsets of the features were tested on GELM neural network and the subset of feature with the highest prediction accuracy is chosen. In addition, the prediction accuracies has presented in Figs. 6. It can be observed that, the prediction accuracy is better for a subset of features compared to when using the full feature set. When the size of dataset is large, the error rate is acceptably in low level. This also confirms that the proposed model retains the characteristic of efficiency even in the case of large dimensionality of the dataset. This nice characteristic is especially important in the big data analysis. But the best accuracy is achieved when the number of features are 20.

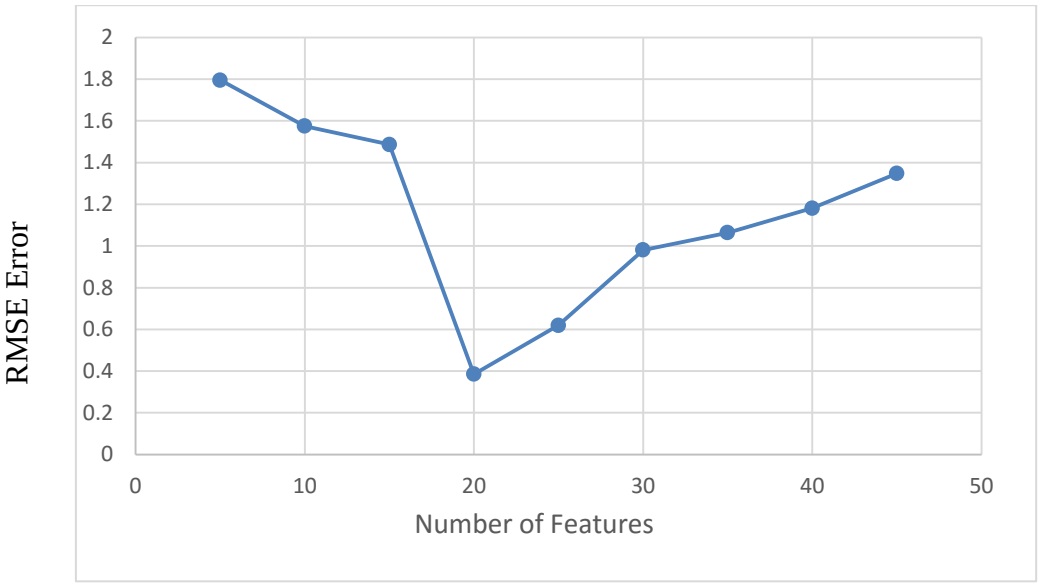

Figure 6. Prediction accuracy with respect to the size of dataset (number of selected features)

Fig. 7 illustrates the computational time of the algorithm in terms of the size of data set. As shown in this figure, the run time increases when the number of features grows. Due to significant reduction of the number of features, better computational efficiency is also achieved.

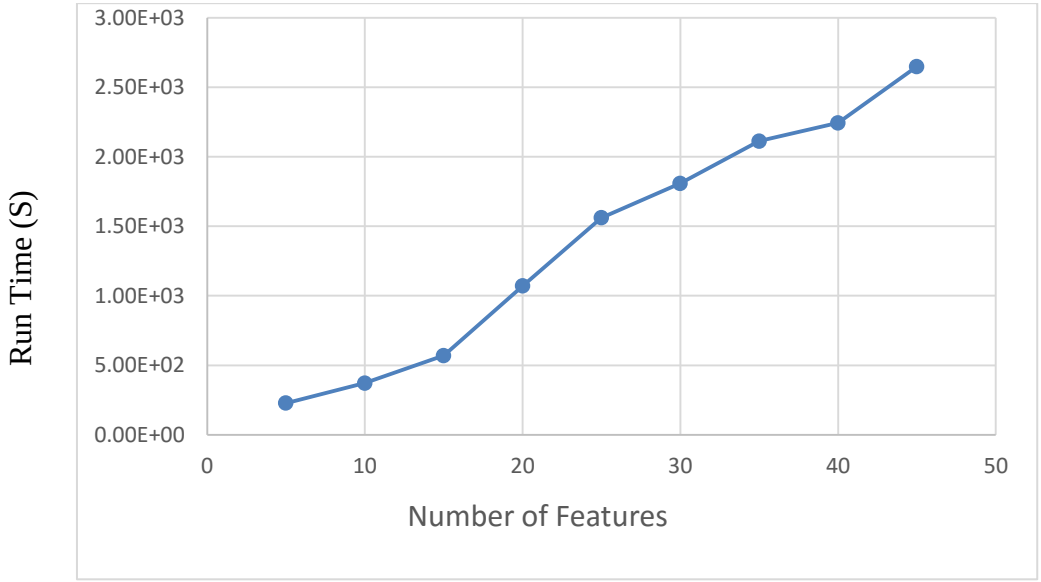

Figure. 7. The relationship between the computational time and the number of features in dataset

Overall, the proposed model selects an optimal subset of features which proved to be superior to other features with regard to prediction of popularity; therefore, it can be considered a variable option for dimensional reduction and modeling online content analysis.

Summary and Conclusion

In this paper, a methodology for modeling and predicting the popularity of online contents was described. This paper investigates the benefit of feature selector on a very popular content-sharing portal – Mashable – by an effective feature selection method based

MI to select a subset of high ranked features, which are strongly relevant but non-redundant for a real-life dataset.

It should be noted that feature selection improves the performance of the model, since the overall performance of our method is excellent in terms of both prediction performance and execution time for this datasets. Using a novel C-MIFS method and a strong ELM neural network, it was shown that the most important predictors of popularity are the time for publishing news (higher number of visitors on weekends) and news topics (lifestyle and social media are the most popular topics on the site).

Possible future works concern the development of a strategy by focusing on different sections of portals (such as how the Mashable "Lifestyle" section differs from the Mashable "Business" section). Also, it would be interesting to learn whether it is possible to predict a Mashable upcoming article popularity by knowing the sharing history of a small number of users.

References

Ahmed, M., Stella Spagna, Huici, F., & Niccolini, S., (2013). A Peek into the Future: Predicting the Evolution of Popularity in User Generated Content 2013. *ACM international conference on web search and data mining*, 607-616.

Balasubramanyan, R., Routledge, B. R., & Smith, N. A., (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.*

Bandari, R., Asur, S., & Huberman, B. A., (2012). The Pulse of News in Social Media: Forecasting Popularity. *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland. ICWSM*.

Battiti, R., (1994) .Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions. Neural Network*, 5 (4), 537–550.

Cao, J., Lin, Z., & Huang, G.B., (2010). Self-adaptive evolutionary extreme learning machine. *Neural Processing Letters*, 36(3) , 285–305.

Cellucci, C. J., Albano, A.M., & Rapp, P. E. (2005). Statistical Validation of Mutual Information Calculations: Comparison of Alternative Numerical Algorithms. *Physical Review*, 71(6), 066208.

Crane, R., & Sornette, D., (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, Moscow, Russia, 15649–15653.

Fernandes, K., Vinagre, P., & Cortez, P. (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In Pereira F., Machado P., Costa E., Cardoso A. (eds). *Progress in Artificial Intelligence.* EPIA 2015. Ledure Notes in Computer Science, Vol. 9273 (535-546). Springer, Heidelberg.

Figueiredo, F., Benevenuto, F., & Almeida, J., (2011). The tube over time: Characterizing popularity growth of YouTube videos. *Proceedings of the fourth ACM international conference on Web search and data mining*.

Foithong, S., Pinngern, O., & Attachoo, B., (2012) .Feature subset selection wrapper based on mutual information and rough sets. *Expert Systems with Applications*, 39, 574–584.

Frenay, B., Doquire, G., & Verleysen, M., (2013).Is mutual information adequate for feature selection in regression?. *Neural Networks*, 48, 1–7.

Hacine-Gharbi, A., Ravier, Ph., Harba, & R., Mohamadi, T., (2012). Low bias histogram-based estimation of mutual information for feature selection. *Pattern Recognition Letters*, 33(10), 1302–1308.

Hensinger, E., Flaounas, I., & Cristianini, N., (2013). Modelling and predicting news popularity", *Pattern Analysis and Application*, 16, 623–635.

Hoque , N., Ahmed, H.A., Bhattacharyya, D.K., & Kalita, J.K., (2016). A Fuzzy Mutual Information-based Feature Selection Method for Classification. *Fuzzy Information and Engineering*, 8(3), 355–384.

Hoque, N., Bhattacharyya, D.K., & Kalita , J.K., (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications*, 41(14), 6371–6385.

Huang, D., & Tommy W.S. Chow (2005). Effective feature selection scheme using mutual information. *Neurocomputing*, 63, 325–342.

Huang, G., Huang, G.B., Song, Sh., & You, K., (2015). Trends in extreme learning machines: A review. *Neural Networks*, 61, 32–48.

Huang, G.B, Zhu, Q.Y, & Siew, Ch. Kh, (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70, 489–501.

Kaltenbrunner, A., Gomez, V., & Lopez, V., (2007). Description and prediction of Slashdot activity. *In: Web Conference, LA-WEB*, 57–66. IEEE, Latin American.

Kim, S.D., Kim, S.H., & Cho, H.G., (2011). Predicting the virtual temperature of weblog articles as a measurement tool for online popularity. Presented In *IEEE 11th International Conference on Computer and Information Technology (CIT)*, 449 –454.

Kwak, H., Lee, C., Park, H., & Moon, S., (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International World Wide Web (WWW) Conference, 26-30.*

Lahoz, D., Lacruz, B., Pedro, M., & Mateo, A., (2013). A multi-objective micro genetic ELM algorithm. *Neuro computing*, 111, 90–103.

Lakkaraju, H., & Ajmera. J., (2011). Attention prediction on social media brand pages. *Proceedings of the 20th ACM international conference on Information and Knowledge Management*, 2157-2160, New York, USA.

Lee, J. G., Moon, S., & Salamatian, K., (2010). An approach to model and predict the popularity of online contents with explanatory factors. Presented in *Web Intelligence and Intelligent Agent Technology*, 623–630.

Lerman, K., & Ghosh, R., (2010). Information contagion: An empirical study of the spread of news on Digg and twitter social networks. Presented in *Proceedings of 4th International Conference on Weblogs and Social Media* (ICWSM-10).

Lerman,K., & Hogg, T., (2010). Using a model of social dynamics to predict popularity of news. *In: Proceedings of the 19th international conference on World Wide Web (WWW)*, 621–630.

Long, X.X., Li, H.D., Fan, W., Xu, Q.S., & Liang, Y.Z., (2013). A model population analysis method for variable selection based on mutual information. *Chemometrics and Intelligent Laboratory Systems*, 121, 75–81.

McCreadie, RMC. Macdonald, C., & Ounis, I., (2010). News article ranking: leveraging the wisdom of bloggers. *In: Proceedings of the 9th international conference on computer-assisted information retrieval (RIAO)*, 40–48

Rezaeenour, J., Yari Eili, M., Roozbahani, Z., & Ebrahimi, M., (2016). Prediction of Protein Thermostability by an Efficient Neural Network Approach. *Journal of Health Management and Informatics, 3(4), 102-110*

Shanon, C.E., (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-428.

Szabo, G., & Huberman, B.A., (2010). Predicting the popularity of online content. *Communications of the ACM*, 53(8), 80–88.

Tatar, A., Antoniadis, P., de Amorim, M.D., & Fdida, S., (2014). From Popularity Prediction to Ranking Online News. *Social Network Analysis and Mining*, 4(1), 4:174.

Tatar, A., de Amorim, M.D, Fdida, S., & Antoniadis, P., (2014). A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1), 1–20.

Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M.D., & et.al (2011). Predicting the popularity of online articles based on user comments. Presented In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS, New York, USA.

Tsagkias, M., Weerkamp, W., & de Rijke. M., (2010). News comments: exploring, modeling, and online prediction. *Lecture Notes in Computer Science springers*, 5993, 191-203.

Tsagkias, M., Weerkamp, W., & Rijke, M.De. (2009). Predicting the volume of comments on online news stories. *In: Proceeding of the 18th ACM conference on Information and knowledge management*, 1765–1768.

Tumasjan, T. O. S., Sandner, P. G., & Welpe, I. M., (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Fourth International AAAI Conference on Weblogs and Social Media*

Vigneswaran, S., Arun Joseph, A., & Rajamanickam, E., (2014). Efficient Analysis of Traffic Accident Using Mining Techniques. *International journal of software and hardware research in engineering*, 2(3), 110-118.

Wang, Zh., Li, M., & Li, J., (2015). A Multi-objective Evolutionary Algorithm for Feature Selection Based on Mutual Information with a New Redundancy Measure. *Information Sciences*, 307, 73-88.

Williams, C., & Gulati, G., (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries. *Annual Meeting of the American Political Science Association*,

Wu, T., Timmers, M., Vleeschauwer, D. D., & Leekwijck, W. V., (2010). On the use of reservoir computing in popularity prediction. *In: Proceedings of International Conference on Evolving Internet, IEEE Computer Society*, Los Alamitos, CA, USA.

Yan, H., Yuan, X.T., Yan, S.C., & Yang, J.Y., (2011). Correntropy based feature selection using binary projection. *Pattern Recognition*, 44(12), 2834–2842.

Yang, J., & Leskovec, J., (2011). Patterns of temporal variation in online media. *Proceedings of the fourth ACM international conference on Web search and data mining*, 177-186.