International Journal of
Applied Mathematics,
Electronics and Computers

IJAMEC

ISSN:2147-8228 www.atscience.org/IJAMEC

Advanced Technology and Science

Original Research Paper

An Effective Classification Method for Facebook Data

Fatih Ertam[1]

Accepted : 12/05/2017 Published: 21/08/2017 **DOI: 10.18100/ijamec.2018Special Issue30463**

Abstract: Today, the use of the internet has become very common. One of the most important reasons for its widespread use is social media tools. Especially Facebook has a very important place in social media tools. For this study, classification was done by using Facebook data. Classifications made by artificial learning algorithms on a previously used data set are compared with accuracy values and learning times. For this purpose, support vector machines (SVM), extreme learning machines (ELM) and K nearest neighbor (kNN) approaches are compared. For the study, SVM and ELM algorithms were observed using different activation functions. For the study with KNN, different K values were tested with different distance metric calculation methods. In the classification approach with ELM, it was observed that higher accuracy values were reached in a shorter time. In addition, Receiver Operating Characteristic (ROC) curves are plotted for the classification in which the best values are obtained for each algorithm.

Keywords: *Extreme learning machine, machine learning, roc curves, social media classification, support vector machine*

1. Introduction

Social media has a great place in today's internet usage. In particular, it is important for companies that want to advertise on social media to invest in which area. Facebook has the highest usage rate among social media tools. The concept of Social Media is at the forefront of many business managers' agenda today [1]. The number of users on the Internet is rapidly increasing day by day. According to Statista of the research sites, the number of internet users will reach 2.95 billion in 2020 [2].

The increase in the usage rates of social media directly increases the usage of internet. Especially popular with social media tools, Facebook maintains its leadership for a long time as its usage rate. According to the data from the same statistics site, in the last quarter of 2016 Facebook reached 1.86 billion monthly active users [3]. Facebook is a social network that aims to connect people with other people and exchange information. Facebook, founded by Harvard University student Mark Zuckerberg on February 4, 2004, was primarily created for Harvard University students. Then Facebook, including schools around Boston, within two months covered the entire Ivy League schools. In the first year; All schools in the United States were available on Facebook. Members priorities were only able to subscribe to the school's e-mail address. Later on in the network and some big companies also joined. In 2006, Facebook opened all e-mail addresses with some age restrictions. Facebook is one of the most visited sites in the world. It is the most visited site in some countries. The site is free for users and receives revenue from banners, logo ads and sponsor groups. The high usage rates of social media have been particularly attractive to companies working in the advertising field. Given its rapid development, it appears that social media brands may become the most important media channel to reach their customers in the near future [4], [5].

[1] *Informatics Department, Firat University, Elazig – 23100, TURKEY*
Email: fatih.ertam@firat.edu.tr
Note: This paper has been presented at the 5th International Conference on Advanced Technology & Sciences (ICAT'17) held in İstanbul (Turkey), May 09-12, 2017.

Using data mining techniques, it is possible to extract the estimated information from the unprocessed data [6]. The social media mining has provided a new head of research. Social media mining integrates social media, social network analysis and data mining to provide a consistent platform for understanding the foundations and potentials of social media mining. It introduces unique problems arising from social media data and provides basic algorithms for data mining and network analysis as well as problems with emerging concepts [7]. In this study, a dataset with Facebook data is classified by SVM, ELM and kNN algorithms using data mining approaches. For the study, the data set in the UCI repository was used [8], [9].

In the second part of this paper, working principles of SVM, ELM and kNN algorithms were mentioned. The third part of the paper mentioned how the data was obtained and which data were used for the classification in experimental studies. In addition, different classification algorithms were compared. In the last part, the results obtained are discussed.

2. Machine Learning Algorithms

In this section, information about SVM, ELM and kNN algorithms has been given from the machine learning methods used for the study.

2.1. Support Vector Machine

SVM is one of the classification techniques based on optimization. It is a method that is mostly used in data mining problems. This method performs the classification with the help of linear function or nonlinear function. It is based on SVM method or the estimation of the most appropriate function to separate the data. This method, which is mostly used with machine learning methods, has started to be preferred in the area of data mining.

SVM learning is a learning algorithm which is used for classification, clustering, density estimation and lastly generating regression results from the data. SVM's theoretical foundations were firstly laid in 1960 by V. Vapnik. It was firstly used in classification in 1995. Vapnik's theory aims to show there is the

solution which comprises the error in the learning clusters and the complexity of the hypothesis space (hypothesis space is expressed according to Vapnik-Chervonenkis (VP) dimension [10].
In general, SVM is designed for two class issues. Besides, many real applications require multi class classification. In order to solve multi class problems with SVM, it is necessary to separate the problem into many two class problems, to train the classifiers to solve the problem and to reconfigure everything for the appropriate output [10]. Multi-class SVMs can be classified under four main categories: One-against-all, one-against-one, Decision Trees based classification and Error-correcting output code ECOC [11]. Different activation functions are used in SVM approach. In this study, linear, quadratic, cubic and Gaussian activation functions were used for classification.

2.2. Extreme Learning Machine

ELM was recommended in Huang et al. [12]. Since 2004, ELM has been using it for Single-hidden Layer Feedforward Neural Networks (SLFNs) training. [13]– [17]. Unlike other feedforward artificial neural networks, ELM does not have to constantly adjust the weight according to the error function. The output weights are determined by calculating the Moore-Penrose generalized inverse matrix while the learning parameters of the neurons in the hidden layer are generated randomly. This is why the speed of education and generalization are high. Fig.-1 shows the general structure of SLFNs.

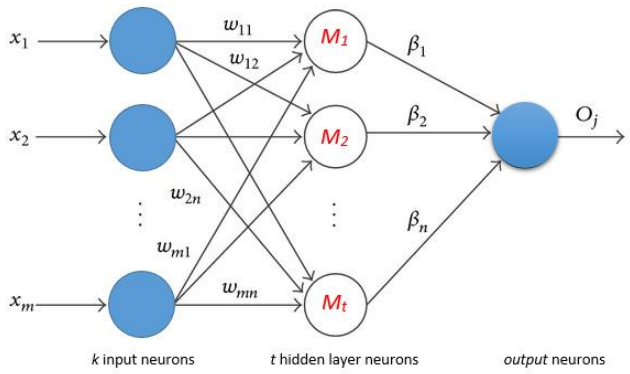


Fig. 1. The network structure of SLFNs.

In this figure, w_{11}, w_{12}, w_{kt} are weights vector connecting the i. hidden neuron. w is weight vector connecting the i. hidden neuron. Given M arbitrary training examples $\{(x_j, r_j)\}_{j=1}^{M} \subset R^d X R^k$, the output of the generalized SLFNs with hidden nodes can be obtained as:

$$f(x_j) = \sum_{i=1}^{t} \beta_i g_i(x_j) \quad (1)$$

$$f(x_j) = \sum_{i=1}^{t} \beta_i G(a_i, b_i, x_j) = o_j, j = 1, \dots, M \quad (2)$$

If the SLFNs can proximate all the M samples without error, that is

$$\sum_{j=1}^{t} \|o_j - r_j\| = 0 \quad (3)$$

There exist pairs of $(a_i, b_i)$and β_i such that:

$$\sum_{i=1}^{t} \beta_i G(a_i, b_i, x_j) = r_j, j = 1, \dots, M \quad (4)$$

The above M equations can also be equivalently expressed in the compact matrix form

$$H\beta = T \quad (5)$$

Where

$$H = \begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_L, b_L, x_1) \\ \vdots & \ddots & \vdots \\ G(a_1, b_1, x_N) & \cdots & G(a_L, b_L, x_M) \end{bmatrix}_{MxL} \quad (6)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{Lxk} \qquad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_M^T \end{bmatrix}_{Mxk} \quad (7)$$

H is named the hidden layer throughtput matrix of the SLFN. β is the output weight matrix. T is the matrix that occur of output labels for the M data pattern. To train a SLFNs as it is said in (2), it is equal to finding the least-square solution $\hat{\beta}$ of linear system (5), that is:

$$\|H\hat{\beta} - T\| = min\beta\|H\beta - T\| \quad (8)$$

On the conditions that the number of hidden nodes L is coequal to the number of different training samples M, it is possible to find a $\hat{\beta}$ which leads to zero training error. The hidden layer output H is a reversible square matrix. The solution of the linear system can thus be given as:

$$\hat{\beta} = H^{-1}T \quad (9)$$

Provided that the number of hidden nodes is L is less than M different training samples to obtain the smallest training error $\|H\beta - T\|$, the answer to the linear system (5) can be obtained in the following way:

$$\hat{\beta} = H^{\dagger}T \quad (10)$$

Where $H^{\dagger}$ is said the Moore-Penrose universalized inverted [18], [20]. Different activation functions are used in ELM approach. In this study, sigmoid, radial basis, tangent sigmoid and hard limit activation functions were used for classification.
It is possible to summarize the operations of the ELM algorithm as follows:

- Assign entry input weight and bias randomly
- Compute the hidden layer output matrix
- Compute the output weight

2.3. K Nearest Neighbors

kNN is an educational and sample-based classification algorithm. Classification of a vector in kNN is done using known vectors. The sample to be tested is processed individually with each sample in the training set. In order to determine the class of the sample to be tested, K samples are selected closest to that sample in the training set. An example of which class is to be tested, if it is the most specific, belongs to this class. Distances between samples are found with distance calculation formulas such as Euclidean, Manhattan, Cosine, Minkowski.
All distance values calculated using the distance metric used are sorted. The least number of K is specified depending on the number of K among the ordered values. K adjacent samples that

are closest to the sample to be tested are determined. The class labels of K neighbors are used for classifying the sample to be tested.

The advantages of this approach are; Applicability is a simple algorithm, it is resistant to noisy educational documents, and is effective if the number of educational patterns is high. Distance based learning algorithm, distance metric selection, high computational cost are disadvantages. For this study, the accuracy ratios were compared by working with different distance calculation metrics and K values.

3. Experimental Studies

The data set used in this study was taken from the UCI repository [8]. There are 500 records in this data set. 40% of the data were used for training and the rest were used for the test. The data set has 19 attributes. The type attribute in the dataset is selected as class.

There are 4 classes that are used on Facebook, including photo, status, video and link. Descriptions of the attributes used are given in Table 1.

Table 1. Features and explanation

No	Explanation
1	Total liked pages
2	Category
3	Month of publication (1-12)
4	The day the submission was published (1-7)
5	Delivery time (1-24)
6	Whether or not the advertisement is paid (1,0)
7	Number of singles who saw
8	Number of clicks and clicks
9	Number of singles who click anywhere in the post
10	Number of people clicking anywhere in the post
11	Any number of clicks in the post
12	Total number of people who like the page
13	The number of people who liked the page afterwards
14	Number of people who liked and joined the page
15	Number of posts in the post
16	Number of likes
17	Number of shares sent
18	Sum of likes, comments and shares
19	Class (Status, Photo, Link, Video)

For this study, a computer with the Matlab 2016b program installed and having 8 GB of RAM and an Intel core i5 2.7 GHZ processor is used. The accuracy value was used to compare the classifiers. Equation 11 is used to find the correctness value.

In Table 2., lines indicate the actual value of the example; and columns of matrix indicate estimated values which were classified or clustered.

Table 2. SVM classification accuracy values and training time

Label	Real Positive	Real Negative
Estimated Positive	True Positive (T_p)	False Positive (F_p)
Estimated Negative	False Negative (F_n)	True Negative (T_n)

The sum of true positive and false positive values divided by the sum of all positive and negative values:

$$Accuracy = \frac{T_p+T_n}{T_p+T_n+F_p+F_n} \quad (11)$$

The normalization procedure shown in Eq. (12) is applied to the data so that the values can range from 0 to 1.

$$Normalization = \frac{x- x_{min}}{x_{max}- x_{min}} \quad (12)$$

Accuracy values obtained with SVM classifications are given in Table 3.

Table 3. SVM classification accuracy values and training time

Activation Function	Accuracy Rate (%)	Time (second)
Linear	91.18	3.58240
Quadratic	91.43	0.30695
Cubic	90.56	0.27425
Gaussian	86.64	0.23885

The best value for the SVM study was found with the quadratic activation function.

The ROC curves of the values obtained with this activation function are shown in Fig. 2.

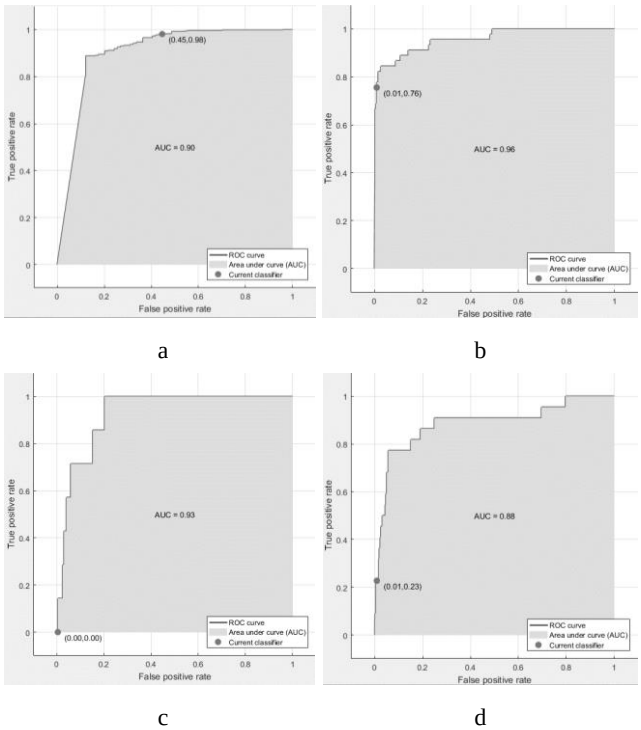


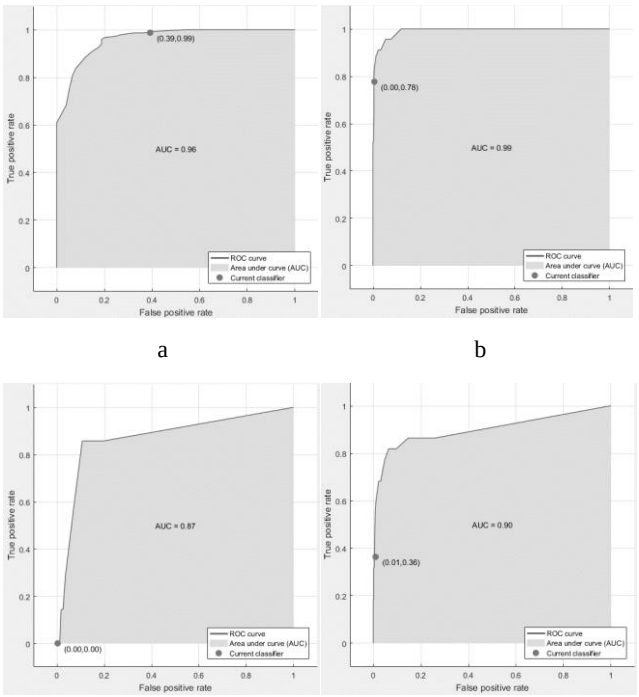
Fig. 2. ROC Curves for SVM. a) Photo, b) Status, c) Video, d) Link.

Accuracy values obtained with ELM classifications are given in Table 4. The number of hidden neurons for the ELM classifier was chosen as 50. The correct rate of classification of the ELM classifier can vary according to different hidden neuron numbers. At the same time, the number of different neurons also influences the duration of the classifier.

Table 4. ELM classification accuracy values and training time

Activation Function	Accuracy Rate (%)	Time (second)
Sigmoid	93.01	0.21454
Radial Basis	92.14	0.15220
Tangent Sigmoid	93.38	0.09140
Hard Limit	89.57	0.98521

In the study with ELM, the best value was obtained by tangent sigmoid activation function.
The ROC curves of the values obtained with this activation function are shown in Fig. 3.



a b

c d

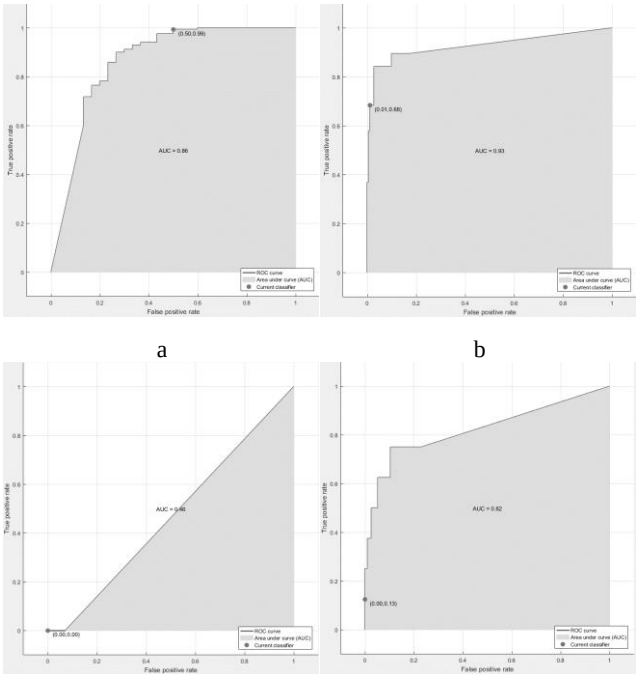
Fig. 3. ROC Curves for ELM. a) Photo, b) Status, c) Video, d) Link.

In the study with kNN, the results obtained by different distance calculation and K values are given in Table 5.

Table 5. kNN classification accuracy values and training time

Distance Metric	K Value	Accuracy Rate (%)	Time (second)
Euclidean	1	90.63	2.00890
Euclidean	10	91.45	0.19928
Euclidean	100	85.03	0.23627
Cosine	10	91.04	0.08121
Minkowski	10	89.11	0.13397

The best value for the study with kNN was found to be the Euclidean distance calculation and the K value of 10. Euclidean, Minkoswki and cosine distance calculations are used in the study with kNN. For the value of K, 1,10 and 100 were chosen. Accordingly, the ROC curves are shown in fig.4.



a b

c d

Fig. 4. ROC Curves for kNN. a) Photo, b) Status, c) Video, d) Link.

4. Conclusions

Social media has a great place in today's internet usage. In particular, it is important for companies that want to advertise on social media to invest in which area. Facebook has the highest usage rate among social media tools. In this study, artificial intelligence techniques were used to classify a data set generated with Facebook data. In order to achieve this purpose, SVM, ELM and kNN are compared.
Studies have been carried out with different activation functions for each classifier. In the SVM study where the quadratic activation function was selected, the accuracy value reached 91.4%. While the lowest value was obtained with the Gaussian function. The ELM classifier using tangent sigmoid activation function has reached 93.4% accuracy value. The lowest value is seen when the hard limit activation function is selected. In the study with kNN, it was observed that the best accuracy value was 91.45%. This accuracy value was obtained in the study where K = 10 was selected and the distance calculation metric was selected as Euclidean. It was observed that the work done with ELM performed faster classification than the work done with SVM. For big data, ELM appears to be a preferred classifier. The results obtained are also shown by ROC curves.

References

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Bus. Horiz., vol. 53, no. 1, pp. 59–68, 2010.

[2] Statista, "Number of Worldwide Social Network Users 2010-2019," Statista, 2016. [Online]. Available: http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[3] Statista, "Facebook users worldwide 2016," statista.com, 2016. [Online]. Available: http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

[4] D. Korschun and S. Du, "How virtual corporate social responsibility

dialogs generate value: A framework and propositions," J. Bus. Res., vol. 66, no. 9, pp. 1494–1504, 2013.
[5] W. G. Mangold and D. J. Faulds, "Social media: The new hybrid element of the promotion mix," Bus. Horiz., vol. 52, no. 4, pp. 357–365, 2009.
[6] E. Turban, R. Sharda, and D. Delen, Decision Support and Business Intelligence Systems, vol. 8th. 2011.
[7] R. Zafarani, M. A. Abbasi, and H. Liu, Social Media Mining. Cambridge: Cambridge University Press, 2014.
[8] U. M. L. Repository, "Facebook metrics Data Set," www.ics.uci.edu, 2016.
[9] S. Moro, P. Rita, and B. Vala, "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach," J. Bus. Res., vol. 69, no. 9, pp. 3341–3351, 2016.
[10] V. N. Vapnik, "Statistical Learning Theory," Interpreting, vol. 2. p. 736, 1998.
[11] A. Statnikov, C. Aliferis, and I. Tsamardinos, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," Bioinformatics, vol. 21, no. 5, pp. 631–643, 2005.
[12] G. Huang, Q. Zhu, and C. Siew, "Extreme Learning Machine : A New Learning Scheme of Feedforward Neural Networks," IEEE Int. Jt. Conf. Neural Networks, vol. 2, pp. 985–990, 2004.
[13] J. Luo, C.-M. Vong, and P.-K. Wong, "Sparse Bayesian extreme learning machine for multi-classification," IEEE Trans. Neural Networks Learn. Syst., vol. 25, no. 4, pp. 836–43, 2014.
[14] G. Bin Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: A survey," Int. J. Mach. Learn. Cybern., vol. 2, no. 2, pp. 107–122, 2011.
[15] F. Ertam and E. Avcı, "A new approach for internet traffic classification: GA-WK-ELM," Measurement, vol. 95, pp. 135–142, 2017.
[16] G. Huang, G. Bin Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," Neural Networks, vol. 61. pp. 32–48, 2015.
[17] F. Ertam and E. Avci, "Classification with Intelligent Systems for Internet Traffic in Enterprise Networks,", Int'l Journal of Computing, Communications & Instrumentation Engg.(IJCCIE) vol. 3, no. 1, pp. 9–15, 2016.
[18] P. Courrieu, "Fast Computation of Moore-Penrose Inverse Matrices," Neural Inf. Process. - Lett. Rev., vol. 8, no. 2, pp. 25–29, 2008.
[19] M.L. Zhang and Z.H. Zhou, ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048, 2007.
[20] F. Ertam and E. Avcı, Network Traffic Classification via Kernel Based Extreme Learning Machine. International Journal of Intelligent Systems and Applications in Engineering, 4 (Special Issue), 109-113, 2016.