International Journal of Engineering & Technology, 7 (4) (2018) 3579-3582

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.19123
Research paper

Gain ratio feed forward neural network algorithm to improve classification accuracy

Helen Josephine V. L [1] *, S. Duraisamy [2]

[1] Research Scholar, Bharathiar University, Coimbatore Department of Computer Applications, CMR Institute of Technology, Bengaluru, India
[2] Department of Computer Science, Chikkanna Govt. Arts College, Tirupur, India
**Corresponding author E-mail: helenjose.cbe@gmail.com*

Abstract

In the field of information technology, there is revolution that has led to an abundance of information in every field through Internet. The rapid growth in the mobile devices indicates that the users and industry are getting more at ease with the mobile environment. An incredible amount of mobile learning systems and users' opinion about these apps are available in the form of reviews on the websites or in the social blogs or feedback. To classify these opinions, Neural Networks algorithm is mostly used to obtain high accuracy. To mine mobile learning app reviews, Gain Ratio based neural network algorithm for opinion mining system is proposed in this research paper. The main focus is to extract the polarity of the reviews, opinion it and conclude whether these reviews are positive or negative or neutral. This research work consists of four steps (i) Estimate score of the words in the review document by using Singular Value Decomposition (SVD) (ii) Feed forward the top ranked words with its weights from the input layer to hidden layer (iii) Calculate gain ratio and select top five positive and negative attributes (iv) Pass the selected attributes from input layer to output layer. This customized neural network classification algorithm helps to improve the classification accuracy.

Keywords: *Classification; Machine Learning; Neural Network; Opinion Mining.*

1. Introduction

One of the techniques to evaluate or mine the feedback and comments of the customers is known as opinion mining. The customer reviews and comments can be gathered from the e-commerce websites, social media, and other blogs. The user comments are classified into two categories, namely positive review comments or negative review comments. Natural Language Processing (NLP) researchers [1] first investigated opinion mining. Automatic text analysis methodology is used for identifying and extracting opinions in a wide array of opinion expressed in the user comments. [2]. The NLP is the most significant technique for the analysis of the sentiments that are applied to the reviews to identify the parts-of-speech. [3]. Moreover, a considerable number of data mining and machine learning algorithms are used to classify the data in the opinion mining. The supervised machine learning methods like Support Vector Machine (SVM), Naïve Bayesian (NB), K-nearest neighbor (KNN), Random Forest (RF) are most extensively used. [4], [5 - 7]. Online Product customer reviews were classified by applying the Sentence weight algorithm [8]. There are several algorithms presented to envisage the polarity of the movie review comments. The sentiment fuzzy classification algorithm was also one among them. The traditional classification method based on parametric multivariate analysis, cluster analysis, and discriminant analysis. These methods are unproductive when data distributions are in nonlinear fashion and even after variable transformation [9].

The neural networks algorithms are played a vital role to classify the positive and negative review comments. The extensive research activities in neural network classification have recognized that neural networks are the potential alternative to various traditional data mining classification methods. The neural networks offer a number of advantages listed below (i) they are data-driven self-adaptive methods which regulate themselves to the data without any explicit specification of distributional or functional form for the model. (ii) Ability to implicitly detect complex nonlinear relationships between independent and dependent variables. (iii) They employ universal functional approximators which helps neural networks to approximate any function with arbitrary accuracy (iv) ability to detect all possible interactions between predictor variables. (v) The availability of multiple training algorithms. [9]

The BPNN (Back propagation neural network) is applied to categorize the customer feedback of the movie and hotel reviews [10]. After analyzing reviews from various researchers, identified that there is a need for an algorithm to improve the accuracy of the text classification. In this research paper, a novel classification method, which is based on the technique of Artificial Neural Networks (ANN), is proposed. The existing neural network algorithm and proposed algorithm classification accuracy are compared with mobile learning app review comments which are obtained from Amazon market. In this proposed work, frequently used words which presented in the review document is also considered along with the opinion word that is presented in the review document. Gain Ratio Feedforward Neural Network (GR-FFNN) algorithm is proposed to improve the accuracy of the classification. This research paper is structured in the following sections. Section 2 describes the existing neural network classifiers. Section 3 explains the proposed neural network classification algorithm GR FFNN. Result and discussion are discussed in Section 4. This paper concludes in Section 5.

2. Neural network classifiers

A standard neural network (NN) contains many simple and connected processors, which are called neurons. Each neuron produces a sequence of real-valued activation. Input neurons get activated through sensors perceiving the environment; other neurons get activated through weighted connections from previously active neurons [11]. In Artificial Neural Network (ANN) the same idea can be imitated by using wires and silicon as living neurons and dendrites. ANNs are consists of multiple nodes, which imitate biological neurons of the human brain. The neurons are connected by links and they interact with each other. The nodes obtain input data and perform simple operations on the data. The result of these operations is passed on to other neurons. The output at each node is entitled as node value. These ANN has been effectively applied to real-world classification problem in the various field of research namely image and pattern recognition, text classification, speech recognition, fault detection, and medical diagnosis and language translation. This section explains the concepts of existing artificial neural network based on classification algorithm.

2.1. LVQ algorithm

The Learning Vector Quantization algorithm (LVQ) is an artificial neural network algorithm which helps in classifying patterns in both binary and multi-class classification. It contains three layers. (i) An input layer (ii) a Kohonen layer which learns and executes the classification (iii) an output layer. The input layer contains one node for each input feature; the Kohonen layer contains equal numbers of nodes for each class; in the output layer, each output node represents a particular class [12]. The training set with the identified classification is given to the neural network along with the initial distribution of the output vector. During training, the output units are set to be positioned to approximate the decision surfaces of the existing Bayesian classifier. One iteration of the training dataset is called an epoch. After all the epochs were completed, an LVQ net is found to classify the input vector by assigning it to the same class as that of the output unit, which has its weight vector very close to the input vector. To reduce the misclassification LVQ classifier is used to adjust boundaries between the categories.

2.2. Elman neural network

An Elman network consists of is a three-layer with the addition of context units. The middle or hidden layer is connected to these context units with fixed weight [13][14]. At every time step, the input is fed-forward and a learning rule is applied. The previous value of the hidden units with context units is saved within the fastened back-connections. At each time step, the input is fed-forward and a learning rule is applied. The previous value of the hidden units in the context units are saved in the fixed back-connections. In other words, the hidden state of each time step is fed to the next time step thereby allowing interactions between time steps in the sequence. Hence, the network maintains an array of state, which permits it to perform such tasks as sequence-prediction. This is the more advanced feature of the network compared to the standard multilayer perceptron. [15]

2.3. Feed forward neural network

McCulloch and Pitts proposed a greatly simplified model of the neuron [16] and these neurons connected together for brain’s activity. The next important growth in artificial neural networks is the concept of a perceptron that was introduced by Frank Rosenblatt in 1958. He stated that Perception has the ability to learn, make decisions, and even translate the languages [17]. This was the foundation of many artificial neural networks. Ivakhnenko ET. Al proposed First generation Multilayer Perceptions. A perceptron calculates a weighted sum of inputs with bias value and special function identified as activation helps to produce output. It propagates inputs by adding all the weighted input and then computing output using sigmoid threshold.

Feed Forward Neural Network (FFNN) are multilayer perceptron with one or many unit layers (hidden) between input and out layers. The simplest MLP structure has one input layer, one hidden layer, and one output layer. Network weights are called Feed Forward as weights flow forward, starting with inputs and with no weights feedback to earlier or current layers [18 - 19]. These types of networks are very powerful and can be extremely complicated. There are no feedback connections in which the outputs of the model are fed back into itself. Fig 1 describes the network.

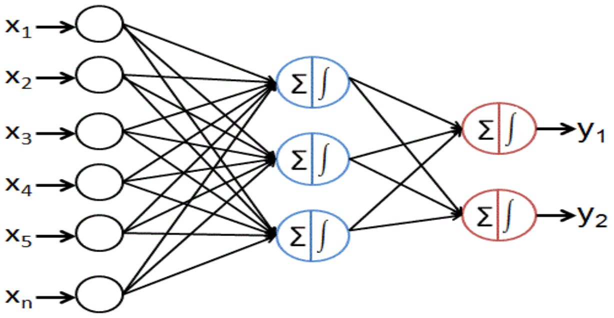

Fig. 1: Multilayer Perception Network.

3. Proposed neural network

This section explained in detail about the proposed work. Through literature review in ANN reveals that ANN algorithm constructed for data classification or pattern recognition is widely apparent to be accurate to the level of 80-90[20]. In order to improve classification accuracy, a neural network based algorithm with well-groomed knowledge is designed. The first and foremost step in opinion mining methods is to extract the opinion words from the opinion text under considerations with the scoring function effectively used by a machine learning algorithm. To obtain classification accuracy both opinion word (feature) extraction and classification algorithm are very important.

The first part of the proposed work emphasis on how to extract opinion word. The opinion words are extracted based on its repeated appearance in all the review documents. The word frequency matrix helps to eliminate common and uncommon words. Singular Value Decomposition algorithm is used to rank the extracted features. The new dataset is obtained and submitted to the machine learning algorithm. The second part of the proposed work concentrates on the classification algorithm. ANN based classification algorithm is proposed to classify the review documents based on their opinion. The proposed algorithm is based on the existing FFNN with two enhancements. (i) Execution of additional weights from the input layer to the output layer. (ii) An improved back propagation algorithm.

The proposed GR FFNN contains three layers, namely the input layer, hidden layer, and output layer. Normally in ANN, all the nodes of the networks are connected with the previous networks. Refer the Fig. 1. But in the proposed GR FFNN, all the nodes of the input layer are not only connected to input layer but also to the output layer. This connection is exactly similar to the connection from the input layer to the hidden layer. The proposed algorithm GRFFNN identifies the top five positive and negative weights using gain ratio method. These additional weights passed to the output layer from the input layer. The summation and activation function will be performed to obtain the output values. Therefore, the proposed GRFFNN is more efficient than FFNN.

The pre-processed dataset x_{one}, x_2, .x_n forms the input for the GR FFNN architecture. The pre-processed dataset helps to improve the classification accuracy in GR FFNN algorithm. Along with the input, random weight w_1, w2...w_n will be fed to the input layer. The product of the input data and the weight will be calculated and presented to the hidden layer in the first epoch. Hidden layer is responsible to perform summation and activation function. The sigmoidal operation is implemented to carry out the activation function. Bias value is set to 0 in the process. A considerable amount of changes will not be seen when the summation of the calculated value with Bias. The output of the hidden layer is the input values to the output layer. Along with these input values, the input layer also provides some more weights to the output layer. In this method, apart from the regular input values, additional weights are feed forward from the input layer to the output layer. The addition, top five positive and top five negative weights are selected with the help of gain ratio.

$$\text{Gain Ratio }(S, A) = \frac{\text{Information Gain }(S,A)}{\text{SplitInformation}(S,A)} \quad (1)$$

Where S represents the sample (training) data, A is the feature. Gain Ratio is a ratio of information gain to the intrinsic information of a split into account [21]. Information gain (IG) measures how much "information" a feature gives us about the class. Features that perfectly partition should give maximal information and unrelated features should give no information. IG also measures the reduction in entropy. Entropy – Entropy: (im)purity in an arbitrary collection of data. The calculation of entropy E(S) for m values in each feature is

$$E(s) = -\sum_{i=1}^{m} p_{i\ \log p_i}\ [22] \quad (2)$$

Where p is the proportion of positive and negative, s is the sample dataset. Entropy value for each class label is calculated. Then, Entropy is calculated for every value in a feature with its corresponding class label. Information Gain (IG) for A feature is calculated using, for the formula [23]

$$\text{Information Gain}(S, A) = E(S)\text{-}\sum_{v \in \text{value } S(A)} \frac{|S_v|}{|S|} E(S_v) \quad (3)$$

Subsequently, every feature with respect to the class labels, Split Information (SI) for A feature is calculated using the following formula [24]

$$\text{SplitInformation}(S, A) = -\sum_{v \in \text{value } S(A)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|} \quad (4)$$

Where E(S) is the entropy of the class label, S is the total number of possibility, S_v is the total number of preferred possibilities and 'v' is the values in a feature [25-26]. From the obtained gain ratio, the top five positive and negative values are chosen for further feeding. The summation and activation function is called to calculate the output values. The process is iterated until the threshold value is obtained or the number of epochs it reached. Parameters used in the GR FFNN algorithm are given in Table 1.

Table 1: Parameters Used in GR FFNN

Parameters	Values
# of neurons in the input layer	60
# of neurons in the hidden layer	30
# of neurons in the output layer	3
# of Hidden layer	1
# of epochs	500
Learning rate	0.1
Momentum	0.5

4. Results and discussions

4.1. Dataset

The mobile learning app reviews have been gathered from Amazon. There are 300 reviews collected randomly from the website. These reviews have been pre-processed. The raw data contains some punctuation marks and some noise. The data cleaning methods help to smooth noisy data, and recognizing removing the outlier. In text mining the basic data cleaning methods like special punctuation removal, stop word elimination, stemming with spell check and lemmatization and other methods have been used. The document contains some stop words, numbers, non-alphabet characters, and vocabularies. Before training the reviews documents are pre-processed by the proposed novel pre-processing techniques. The newly proposed data set is divided into two sets which are the training set and the testing set. In the training set, there are 120 input vectors with 453 and features and there are 180 input vectors with 453 features in the testing set.

Implementation is done in python using Anaconda. Anaconda is a package which provides python and most of the libraries which used for machine learning pre-installed. Scikit learn library has been used which contains a variety of machine learning library for the Python programming language.

4.2. Experimental dataset

This pre-processed dataset is fed in GR FFNN for classification. The classification accuracy achieved from the experiment is 85.33%. The experiments were conducted with the existing ANN classifiers namely, LVQ, Elman, and FFNN. The classification accuracy and other classification metrics namely precision, recall, F-Measure has evaluated, and the values were compared. The Proposed GR FFNN method achieved the classification accuracy of 85.33%, whereas the other ANN classifiers LVQ gained 60.67%, Elman neural network obtained 79.67% and FFNN gained 82.33%. The classification accuracies are tabulated in Table II. The classification accuracy obtained from the GR FFNN is 3% more than the existing FFNN. The classification accuracy has been improved because of the additional weight. The top five positive and negative features obtained by gain ratio formula fed to the output layer from the input layer.

Table 2: Comparing Classification Accuracy of GR FFNN with Existing Neural Network Classifiers

Algorithm	Classification Accuracy %
LVQ	60.67
Elman	79.67
FFNN	82.33
Proposed GR FFNN	85.33

The comparative analysis it is evident to note that the proposed GR FFNN classifier supersedes the existing ANN classifiers. The results are shown graphically in Fig. 2

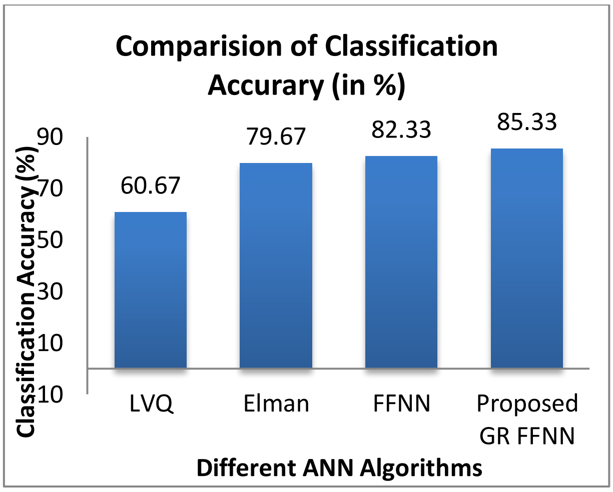

Fig. 2: Comparison of GR FFNN with Existing.

Neural Network Algorithms

The classification accuracy is validated using Precision, Recall, and F-Measure. Precision reveals the probability that the retrieved document is relevant. Recall shows the probability that a relevant document in a search. Whereas, F-Measure is the harmonic mean of Precision and Recall. The validated measures precision, recall, and f-measure are shown in Table 3 for the FFNN and the proposed GR FFNN classifiers.

The measure of validation namely precision, recall and f-measure are represented graphically in Fig.3. This shows that an improvement in the accuracy rate.

Table 3: Precision, Recall and F-Measure of mobile Learning App Review Dataset

Algorithm	Precision	Recall	F-Measure
LVQ	0.62	0.61	0.64
Elman	0.81	0.80	0.76
FFNN	0.83	0.82	0.82
Proposed GR FFNN	0.85	0.86	0.85

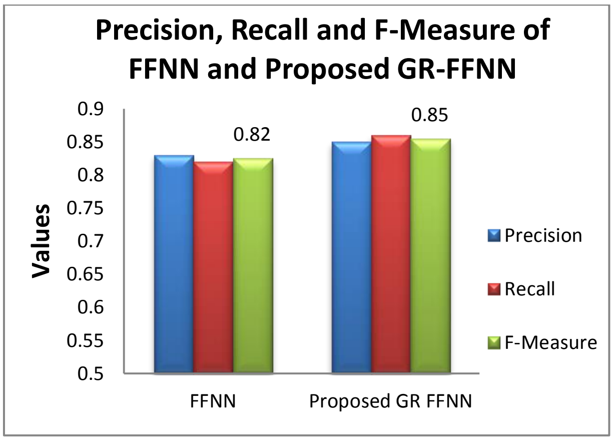

Fig. 3: Precision, Recall, and F –Measure of Mobile Learning App Review Dataset.

Thus proposed algorithm GR FFNN is validated against the existing ANN classifiers with the classification accuracy measures Precision, Recall and F-Measure. This result reveals that the proposed algorithm yields a better result than the existing ANN algorithm.

5. Conclusion

In this paper, an enhanced neural network classification algorithm has been proposed to classify the opinions of learners of the mobile learning system. The classification accuracy for opinion has been evaluated using the proposed GR FFNN algorithm. The feature was extracted from the dataset and ranked using Gain Ratio. The obtained features were employed in the GR FFNN algorithm. In order to substantiate the proposed algorithm's performance, ANN based classifiers have been discussed. The proposed methodology includes the architecture and algorithm. The important intervention in the summation and activation function that has brought a difference in the performance has been explained. The higher point and lower point of the values, calculated through Gain Ratio, have brought the significant improvement in the classification accuracy. The experimental results have shown that high accuracy gained in GR FFNN algorithm over the other existing ANN algorithms.

References

[1] Dave, D., Lawrence A., and Pennock D. Mining the Peanut Gallery Opinion Extraction and Semantic Classification of Product Reviews, Proceedings of International World Wide Web Conference (WWW ’03), 2003.

[2] Talemura Junichi, “Virtual reviews for collaborative exploration of movie reviews”, in proceedings of Intelligent User Interfaces (IUI) pages 272-275, 2000.

[3] S. Thanangthanakij, E. Pacharawongsakda, N. Tongtep, P. Aimmanee, T. Theeramunkong, “An Empirical Study on Multi- Dimensional Sentiment Analysis from User Service Reviews”, Knowledge, Information and Creativity Support Systems, pp. 58 – 65, IEEE, 2012.

[4] C. Zhang, W. Zuo, T. Peng and F. He, “Sentiment Classification Reviews Using Machine Learning Methods Based on String ernel”, Convergence and Hybrid Information Technology, Vol. 2, pp. 909 – 914, IEEE, 2008.

[5] A.Khan, B.Baharudin, K.khan, “Sentence Based Sentiment Classification from Online Customer Reviews”, Frontiers of information Technology, ACM, 2010.

[6] N.Aleebrahim, M.Fathian and M.Reza Gholamian, “Sentiment Classification of Online Product Reviews Using Product Features”, Data Mining and Intelligent Information Technology Applications, pp. 242 – 245, IEEE, 2010.

[7] K. Gayathri, A. Marimuthu, “Text Document Pre-Processing with the KNN for Classification Using the SVM”, Intelligent Systems and Control, pp. 453 – 457, IEEE, 2012.

[8] X.Hu and B.Wu, “Classification and Summarization of Pros and Cons for Customer Reviews”, Web Intelligence and Intelligent Agent Technologies, Vol. 3, pp. 73 – 76, IEEE, 2009.

[9] Jack V.Tu et.al Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, Journal of Clinical Epidemiology, Volume 49, Issue 11, November 1996, Pages 1225-1231, https://doi.org/10.1016/S0895-4356(96)00002-9.

[10] A.Sharma and S.Dey, “A Document-Level Sentiment Analysis Approach Using Artificial Neural Network and Sentiment Lexicons”, ACM SIGAPP Applied Computing Review, VOL. 12, pp. 67-75, ACM, 2012.

[11] Jurgen Schmidhuber et.al, Deep Learning in Neural Networks: An Overview (NPTL)

[12] Yang, D, Chen, G., Wang, and H. et al. Learning vector quantization neural network method for network intrusion detection Wuhan Univ. J. of Nat. Sci. (2007) 12: 147. https://doi.org/10.1007/s11859-006-0258-z.

[13] Elman, Jeffrey L. (1990).”Finding structure in time." Cognitive Science, 14, pp. 179-211. https://doi.org/10.1207/s15516709cog1402_1.

[14] Fausett, Laurene. (1994). Fundamentals of neural networks: Architectures, algorithms, and applications. New Jersey: Prentice Hall.

[15] Cruse, Holk; Neural Networks as Cybernetic Systems, 2nd and revised edition

[16] McCulloch, W.Pitts, W. “A logical calculus of the ideas immanent in nervous activity”, The Bulletin of Mathematical Biophysics, Volume 5, pp.115-133, 1943. https://doi.org/10.1007/BF02478259.

[17] Rosenblatt, Frant, “The Perceptron: A probabilistic Model for Information Storage and Organization in the Brain”, Cornell Aeronautical Laboratory, Psychological Review, volume 65, No. 6, pp. 386-408, 1958.

[18] Mu-Song Chen, M.T. Manry, “Power Series Analyses of Back-Propagation Neural Networks”, International Joint Conference on Neural Networks, Volume I, pp 295-300, 1991.

[19] Mu-Song Chen, M.T. Manry, “Nonlinear Modeling of Back-Propagation Neural Networks”, International Joint Conference on Neural Networks, Volume II, pp 899, 1991.

[20] Peter D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 471-424, 2002.

[21] http://www.saedsayad.com/decision_tree.htm.

[22] Shikha Chourasia “Survey paper on improved methods of ID3 decision tree classification” International Journal of Scientific and Research Publications, Volume 3, Issue 12, December 2013

[23] Muharram M.A., Smith G.D. (2004) Evolutionary Feature Construction Using Information Gain and Gini Index. In: Keijzer M., O’Reilly UM., Lucas S., Costa E., Soule T. (eds) Genetic Programming. EuroGP 2004. Lecture Notes in Computer Science, vol 3003. Springer, Berlin, Heidelberg.

[24] Decision Tree Algorithm. In: Khachidze V., Wang T., Siddiqui S., Liu V., Cappuccio S., Lim A. (eds) Contemporary Research on E-business Technology and Strategy. iCETS 2012. Communications in Computer and Information Science, vol 332. Springer, Berlin, Heidelberg.

[25] He Zhang, Runjing Zhou, The analysis and optimization of decision tree based on ID3 algorithm, 2017 9th International Conference on Modelling, Identification and Control (ICMIC).