Combining support vector machine with radial basis function kernel and information gain for sentiment analysis of movie reviews

Z Abidin*, W Destian1, and R Umer2
1 Department of Computer Science, Universitas Negeri Semarang, Indonesia
2 Department of Computer Science, Balochistan University of Information Technology Engineering and Management Sciences, Quetta, Pakistan

Abstract. A movie review can be considered as an essential data source for both movie producers and potential consumers. The reviews are usually used as a benchmark for knowing the movie's quality and are utilized as references to whether the film is worth watching. The study of sentiment analysis has received more attention from researchers. This current study aims to classify the sentiment analysis of movie reviews obtained from the IMDb site. The support vector machine (SVM) method was employed to classify the movie review's sentiments. Meanwhile, the radial basis function (RBF) kernel and information gain (IG) were used to enhance classification. Feature selection was conducted by removing irrelevant features and selecting features with a strong correlation to classification. The IG algorithm was used for feature selection. The current study revealed that the classification accuracy of movie review sentiment analysis using the SVM algorithm, SVM + RBF kernel, and SVM+RBF Kernel+IG are 81.50%, 82.25%, and 87.25%, respectively.

1. Introduction

Research on sentiment analysis has gained more attention from researchers and business owners in marketing their products or services. Business owners can find out how the market responds to their products or services through customers' reviews, ratings, advertisements, news, and comments from electronic media [1]. Sentiment analysis can be described as building a computational system by gathering opinions about an object from various sources and analyzing it [2]. According to Chaturvedi et al. [3], sentiment analysis aims to determine whether the dataset contents are in the form of text, are positive, negative, or neutral opinions. Sentiment analysis has been using in various aspects of researches, such as ranging from sales predictions [4], politics [5], and taking investor decisions [6].

Consumer opinion is now becoming one of the most important sources of information on various products including in the film industry [7]. One of which is the Internet Movie Database (IMDb) site, a site that deals with film and film production. Various classification algorithms for analyzing product review sentiments have been proposed by researchers, such as Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN) [8], Maximum Entropy (ME) [9], Artificial Neural Network (ANN) [10], and Multilayer Layer Perceptron (MLP) [11].

SVM algorithm has advantages compared to other algorithms because it works well on many feature dimensions and has overfitting protection. It means that the SVM algorithm's use does not always depend on the number of dataset features [12]. Bhavitha et al. [12] comparing SVM, and KNN algorithms to the classification of movie review sentiment analysis, the results SVM managed to obtain.
the best accuracy results. Research by Shivaprasad and Shetty [13] comparing SVM, NB, and ME algorithms to conduct sentiment analysis on data reviews on online store sites, the result is that the SVM algorithm has obtained the highest average accuracy.

According to Gomes, Prudêncio, Soares, Rossi and Carvalho [14], SVM performance is very dependent on kernel selection. RBF kernel is considered can produce better classification performance when appropriate gamma parameter values and soft margin constant C are selected [15]. Ahmed et al. [16] compare the SVM poly kernel algorithm, SVM RBF kernel, NB, and MLP to classify movie review sentiment analysis. RBF kernel SVM gets the best accuracy results. Jadav and Vaghela [17] classify sentiment analysis on the polarity movie review dataset, twitter dataset, and gold dataset from amazon.com. They used the SVM RBF kernel to obtain the best accuracy.

Another problem with the classification of sentiment analysis is the number of features used in a dataset [18]. According to Wang et al. [18], feature selection has the concept of reducing a large feature space by removing features that are less relevant in classification. Information gain (IG) is considered an appropriate method to deal with gaining features based on data rank (top k) [19]. Chandani and Wahono [10] compare the SVM, NB, ANN algorithm in the movie review sentiment analysis with a combination of feature selection, namely IG, chi-square, forward selection, and backward selection. The result is IG get the best average accuracy improvement. Kalaivani and Shunnuganathan [1] compare the SVM, NB, and KNN classification algorithms in the classification of movie review sentiment analysis with the application of IG feature selection. Their study results indicated that the combination of SVM and IG successfully obtained the best accuracy.

Based on the description of the problem above, this current study aims to investigate the use of the SVM algorithm, with the RBF kernel application and IG feature selection towards movie review sentiment analysis.

2. Methods

The classification process of sentiment analysis in the current study consisted of several main stages: text preprocessing, word vectorization, feature selection, and sentiment analysis classification. The data used in this study is the Data Movie Review Polarity Dataset V2.0 obtained from http://www.cs.cornell.edu/people/pabo/-movie-review-data. The movie review dataset comprises 2000 movie review data with 1000 reviews labelled positive, and 1000 reviews labelled negative. Data were split into two sets: training data and testing data, with a percentage of 80% (N=1600 training data) and 20% (N=400 testing data), respectively.

Before analyzing the text reviews, the text preprocessing stage was conducted by changing unstructured data into structured data. The text preprocessing consisted of tokenizing, filter token, stopword filtering, and stemming. Word vectorization was also performed by converting a collection of tokens/terms from bag-of-words in a dataset to numeric vector features. This process was carried out to change the text form of the dataset into numeric vector features. The method used for word vectorization in this current study is TF-IDF. The numerical vector values yielded in this technique are called weights. The weight for each token/term can be calculated by using equation 1.

\[ idf_k = \log \frac{N}{n_k} \]

The TF-IDF value can be obtained by using equation 2.

\[ tfidf(t, d) = tf(t, d) * idf_t \]
2.1. Information Gain (IG) Feature Selection

IG feature selection method used in this current study aims to reduce the number of too large features. Only features with a strong correlation to classification were selected, while the less relevant features were eliminated. The IG feature selection can be performed by calculating the entropy of each feature in the dataset (see equation 3).

\[
info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)
\]

(3)

Then calculate the information value of the two samples using equation 4.

\[
info_A(D) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} \cdot info(D_j)
\]

(4)

The IG value for each feature can then be obtained using equation 5.

\[
Gain(A) = info(D) - info_A(D)
\]

(5)

Finally, sort the ranking of features based on the IG value, from the highest to lowest feature values, and the selected features are ready to be used for the classification process.

2.2. Sentiment Analysis Classification

At this stage, the SVM algorithm with the RBF kernel was used to classify the movie review dataset, which has previously been through text preprocessing, word vectorization, and feature selection. The training data were grouped into certain classes and then determined the limit (margin). When connecting every point in a class from one data set to every other data set in that class, a decision boundary line will appear defining this class boundary. Next, determine the best hyperplane with maximum margins. The hyperplane is determined by using equation 6.

\[
w \cdot x + b = 0
\]

(6)

Controlling the data labelling errors can be done using the concept of soft margins by using equation 7. Parameter \( C \) function used as an optimization controller between margin and \( t_i \) misclassification.

\[
\min_{w,b,t} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{m} t_i
\]

(7)

Finally, obtaining the kernel value using the RBF kernel function, as shown in equation 8. In the RBF kernel function, parameter \( \gamma \) (gamma) is used as a learning speed controller of the SVM algorithm.

\[
K(x,x_k) = \exp\left(-\frac{||x-x_k||^2}{2\sigma^2}\right)
\]

(8)

The labelling process can be performed by using the separator function in equation 9. If the result is +1, then the data are categorized into a positive class; otherwise, the data are classified into a negative class.

\[
f(x) = \gamma \ a'K(x_{test},x_{train}) + b
\]

(9)

3. Results and Discussion

The current study employed the RBF kernel and IG feature selection to improve the movie review sentiment analysis accuracy. The classification algorithm used in this study was the SVM algorithm. This study compares the SVM algorithm, the SVM algorithm with the RBF kernel, and the SVM algorithm with the RBF kernel and IG. The higher the accuracy results obtained, the better the algorithm is rated.
The RBF kernel was applied to the SVM by initializing parameter values based on experiments. The accuracy results of the variation of the $C$ parameter are shown in figure 1. Further, the accuracy results of the variation of the $gamma$ parameter can be seen in figure 2. Based on the experiment in varying the $C$ and $gamma$ parameter, the authors initialized parameters $C = 10$ and $gamma = 1.5$ in the SVM algorithm RBF kernel and resulted in an accuracy rate of 82.25%.

![Figure 1. Accuracy result from the experiment $C$ parameters](image1)

![Figure 2. Accuracy result from the experiment $gamma$ parameters](image2)

![Figure 3. The results of the accuracy of three different combined methods](image3)
Figure 3 shows that the combination of the SVM algorithm with the RBF kernel has resulted in increased accuracy by 0.75%, from 81.5% to 82.25%. Moreover, when the IG was applied in SVM+RBF Kernel, the accuracy increased by 5%, from 82.25% to 87.25%.

Table 1. Comparison of accuracy results

| Author                     | Method                                      | Best result |
|----------------------------|---------------------------------------------|-------------|
| Chandani and Wahono [10]   | SVM                                         | 81,10%      |
|                            | NB                                          | 74,00%      |
|                            | ANN                                         | 51,80%      |
| Kalaivani and Shunmuganathan [1] | SVM + n-gram + IG (top k=500) | 81,45%      |
|                            | NB + n-gram + IG (top k=500)               | 75,55%      |
|                            | KNN + n-gram + IG (top k=500)              | 68,7%       |
| Proposed method            | SVM RBF kernel + IG (top k = 2100)         | 87,25%      |

Table 1 presents the comparison study results from previous studies with its methods used. All the authors were using the same dataset. It can be seen from Table 1 that the proposed method (SVM + RBF kernel+IG (top k = 2100) has provided improved results.

4. Conclusions and Limitations
The current study has examined movie reviews with the notion of sentiment analysis. The SVM algorithm, RBF kernel, and IG have provided evidence that the accuracy results obtained from the combination of these methods yield better improvement. The RBF kernel and IG use in this current study have proven the classification process's enhancements. Modification of C and gamma parameters also affects the sentiment analysis results.

References
[1] Kalaivani P and Shunmuganathan K L 2013 Indian J. Compute. Sci. Eng. 4 285
[2] Medhat W, Hassan A and Korashy H 2014 Ain Shams Eng. J. 5 1093
[3] Chaturvedi I, Satapathy R, Cavallari S and Cambria E 2019 Pattern Recognit. Lett. 125 264
[4] Fan Z-P, Che Y-J and Chen Z-Y 2017 J. Bus. Res. 74 90
[5] Park S J, Lim Y S, Sams S, Nam S M and Park H W 2011 Soc. Sci. Compute. Rev. 29 288
[6] Dergiades T 2012 Econ. Lett. 116 404
[7] Zhu J, Wang H, Zhu M, Tsou B K and Ma M 2011 IEEE Trans. Affec. Comput. 2 37
[8] Maheswari K 2018 Int. J. Res. Anal. Rev 5 422
[9] Xie X, Ge S, Hu F, Xie M and Jiang N 2019 Soft Comput. 23 599
[10] Chandani V and Wahono R S 2015 J. Intel. Syst. 1 56
[11] Ahmad M, Aftab S, Muhammad S S and Ahmad S 2017 Int. J. Multidiscip. Sci. Eng 8 27
[12] Bhavitha B K, Rodrigues A P and Chiplunkar N N 2017 IEEE Xplore 17042250 216
[13] Shivaprasad T K and Shetty J 2017 IEEE Xplore 17042231 298
[14] Gomes T A F, Prudêncio R B C, Soares C, Rossi A L D and Carvalho A 2012 Neurocomputing 75 3
[15] Zhou S-S, Liu H-W and Ye F 2009 Neurocomputing 72 2931
[16] Ahmed E, Sazzad M A U, Islam M T, Azad M, Islam S and Ali M H 2017 IEEE Xplore 17261520 86
[17] Jadav B M and Vaghela V B 2016 Int. J. Compute. Appl. 146 26
[18] Wang S, Li D, Song X, Wei Y and Li H 2011 Expert Syst. Appl. 38 8696
[19] Aggarwal C C and Philip S Y 2008 Privacy-preserving data mining: Models and algorithms (US: Springer-Verlag)