Evolution of hybrid distance based kNN classification

N. Suresh Kumar, Pothina Praveena
Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM Deemed to be University, Visakhapatnam, India

Article Info

Article history:

Received Jun 8, 2020
Revised Feb 7, 2021
Accepted Mar 24, 2021

Keywords:

Artificial bee colony
Classification
Distance weighted
K-nearest neighbour
Opinion mining

ABSTRACT

The evolution of classification of opinion mining and user review analysis span from decades reaching into ubiquitous computing in efforts such as movie review analysis. The performance of linear and non-linear models are discussed to classify the positive and negative reviews of movie data sets. The effectiveness of linear and non-linear algorithms are tested and compared in-terms of average accuracy. The performance of various algorithms is tested by implementing them on internet movie data base (IMDB). The hybrid kNN model optimizes the performance classification interns of accuracy. The accuracy of polarity prediction rate is improved with random-distance-weighted-kNN-ABC when compared with kNN algorithm applied alone.

Corresponding Author:

N. Suresh Kumar
Department of Computer Science and Engineering
GITAM Institute of Technology, GITAM Deemed to be University
Visakhapatnam, India
Email: nskgitam2009@gmail.com

1. INTRODUCTION

The classification techniques plays very important role in obtaining best fit data set. In the present paper various classification techniques are presented with their merits and demerits. The algorithms are categorised based on the applications in which they are used [1], [2]. The classification is carried by evaluating user “opinion” on a particular movie.

Opinion mining is one of the important and exquisite topics in recent days conferred in many types of data analytics. The online data analysis plays very important role in future prediction such as politics, movie reviews, sometimes it will also dictates the market strategies. Many organizations, political parties, news channels and film industries are very much fascinated to know about customer’s or user’s perception about their product or their services. The customer’s reviews are very important for the organizations to evaluate the companies’ services or products. It is very important for the companies or political parties to predict the plan of action in near future based on user’s reviews.

The companies, or movie directors, or political parties gain insight knowledge which can help them to reach the target in the market or society. The analysis of user reviews and their opinion will help the product developers or service providers [3] in terms of performance enhancement. For instance the movies reviews are posted by the netizens are viewed by many other users and do their comments also. This will help the other users to make a verdict about the film they are going to see or to get an opinion on already released movie. These reviews also help the film casting section to understand the people cognizance set.

The predictive models are very important for organizations to retrieve public opinions for captivating wise decisions. It is very difficult and multifaceted to processes huge data available online through internet. Variety of data contains diversified features and may causes unnecessary confusion with existing models. The confusion may lead poor accuracy in retrieval of required data. In all these circumstances users are allowed to expresses their reviews or feedback in their natural language.

Hence it is very important and complex task for the computer to analyse the natural language to get a more accurate opinion from large database. Natural language processing (NLP) is one of the tools applied on these reviews to retrieve an opinion from the reviews. The accuracy of prediction is measured at each algorithm. Generally NLP involves several steps like word tokenization, labelling, lemmatizing, parsing, stemming, and data analysis of data sets to perform sentiment analysis. In the recent days it became easier to analyse the natural language when compared to past with the help of NLTK library available in PyCharm. The user reviews collected from unprocessed open source data base called internet movie data base (IMDB) [4] which is first pre-processed and then classified based on the features defined. The use-full data or data patterns can be extracted by implementing various advanced data mining techniques. It is found that kNN algorithm is one of the best method used to classify when an unknown instance is presented to the supervised learning system and regression models. In the present paper, k-nearest neighbour (kNN) classification technique is used with artificial bee colony (ABC) to optimize the performance of kNN classifier.

2. EXISTING TECHNIQUE

In general classification plays very important role in categorising and predicting the group of data element with the involvement of artificial intelligence. Classification techniques are pulled into various extents such as data retrieval, signal processing, bio-informatics, robotics, and psychology [5]. The supervised learning system is used to gather and derive the purposeful dataset from the trained dataset. The training dataset contains the vector quantity and desired purposeful output [6], [7]. Different supervised machine learning algorithms in support of current work are:

a. Linear algorithms
- Logistic regression (LR) [8], [9]
- Linear discriminant analysis (LDA) [10]-[12]

b. Non-linear algorithms
- Naïve-Bayes classifier (NB) [2], [13]-[16]
- Support vector machine classifier (SVM) [13], [17]-[19]
- k-nearest neighbours (KNN) [13], [20]-[22]

2.1. Linear algorithms

2.1.1. Logistic regression classifier

Schmidt *et al.* [8] and Vanashri [9] has discussed key points of logistic regression on movie review analysis. It is linear model which is also called as log-linear classifier or maximum-entropy (ME) classifier. A logistic function is used for deducing outcome. The cost function is minimized with L2 regularization. The problem of optimization solved with L1 regularization [8], [9]. It is proved that it produced best accuracy results when compared with query expansion. Table 1 depicts the comparisons with existing techniques.

2.1.2. Linear discriminant analysis (LDA)

Andrew [10] proposed that it outperforms than LRC. It is proved that the LDA produces 67% of accuracy on standard bench mark IMDB dataset. The logistic regression algorithm has limitation of analysis for two sample classification. This can be overcome by LDA analysis. When samples are separated the logistic regression becomes unstable. For single variable the mean and covariance parameters are measured at each sample. Similarly mean and covariance matrix is calculated for multi-variant input. The discriminant analysis is used in the case of categorization type datasets. The discriminant analysis is used when categorization is not preferred and interested odd relations of each snippet. LDA is used when tries to separate two or more events or classes. This type of linear classifier is used for dimension reduction. The LDA is mostly used for each observation of measurement of independent variable [11], [12].

2.2. Non-linear algorithms

2.2.1. Naïve-bayes classifier (NB)

Padmavathi *et al.* [2], Adebayo [13], and Palk *et al.* [14] has discussed the merits of NB classifier and compared the results with other non-linear algorithms. It is proved that NB classifer able to produce good results when compared with LR and LDA algorithms. With NB classifier almost 75.5% of accuracy is achieved from user move reviews. The Bayes theorem is implemented on the feature pairs in this classification technique. Assume that there is a class variable ‘B’ and A_1 to A_n are dependent vector features assumed through maximum likelihood, and hence the bayes is derived as shown in (1) [15], [16].

$$P(B\left|A_1,\dots,A_n\right) = \frac{P(B)P(A_1,\dots,A_n|B)}{P(A_1,\dots,A_n)} \quad (1)$$

and, from the theory of independence it is derived as shown in (2).

$$P(A_i|B,A_1,\dots,A_{i-1},A_{i+1},\dots,A_n) \quad (2)$$

For every 'i' this can be written as,

$$P(B\left|A_1,\dots,A_n\right) = \frac{P(B)\pi_{i=1}^{n}P(A_1|B)}{P(A_1,\dots,A_n)}$$

Therefor the classification techniques can be applied with classification as,

$$P(B\mid A_1,\dots,A_n)\alpha P(B)\prod_{i=1}^{n}P(A_i\mid B)$$

$$\Downarrow$$

$$\hat{B} = \arg\max_{B} P(B)\prod_{i=1}^{n}P(A_i\mid B)$$

where, *P(B|A)* is the probability of subsequent target B for a given attribute A.
P(B) is the probability of a class of values.
P(A|B) is the probability of a predictor of a given target
P(A) is the past probability of occurring of the attributes.

The confusing factors may cause uncertainty in record prediction. This method isolates the noise factors and irrelevant attributes while assessing conditional probabilities.

2.2.2. SVM classifier

Adebayo [13] has analysed movie review dataset with SVM classification and proved that SVM is superior to NB classifier with subjective feature extraction. It is shown that almost 80.5% accuracy is achieved with SVM classification technique. SVM was first introduced by Cortes and Vapnik and modified later [17], [18]. The SVM mostly used on multi-dimensional space. Hence SVM has the capability of multi-class classification. SVM uses input space for computing kernels [19]. Here every data set represented as vector categorised as class. In the sequence of procedure the margin between classes is measured. The class margin helps to reduce uncertain decisions.

2.2.3. K-nearest neighbours (KNN)

Dudani [20] and Halil [21] presented different techniques along with kNN classification. Distance weighted kNN produced 92% accuracy in opinion evaluation of the user movie reviews. The KNN algorithm saves all newly classified and existing snippets based on distance function. The statistical estimation of prediction analysis can be done with KNN algorithm with non-parametric estimation technique. The word snippets are considered by assigning ranks to the specified snippet. The rank is assigned by majority of the neighbouring data classification. In the rank based analysis kNN is recognised as best suitable algorithm for data analysis. The nearest neighbours are classified and analysed with the help distance function [22]. Hence forth the datasets are labelled and trained. The hamming distance is used in the definite variable case of dataset. In the case of word analysis if the corpus data word is nearest to the positive or negative snippet in the dataset then the distance is '0' otherwise the distance is '1'. The KNN is one of the prominent algorithms used when there is no idea or little knowledge about the dataset to be classified. In the present paper the kNN algorithm also tested with optimization technique called artificial bee colony (ABC) to enhance the prediction rate. The unknown datasets are classified with reference to different numbers of '*k*- neighbours'. The population increase as the neighbours space increase. This may lead complexity in computation which further has impact on accuracy. This problem is overcome with the help of ABC algorithm which is discussed in section 4.

- Hamming distance

The kNN is a non-parametric algorithm and stores all the existing circumstances or cases and accordingly derives a new class of case. The new class of cases are purely depends on precise distance measured between k-neighbour data points. The distance between two data points is measured by similarity measure or distance measure. There are several distance functions effectively involved in deriving a class of a case. Some of the functions are Euclidean, Manhattan, and Minkowski distance functions [23]. But all the three distances are used in continuous variable classification. Euclidean distance represented as,

$$d(x, y) = \sqrt{\frac{\sum_{i=1}^{k}(x_i - y_i)^2}{k}} \tag{3}$$

where, x and y are points are a line segment. Such as, $x=(x_1, x_2,..., x_i)$ and $y=(y_1, y_2, ..., y_i)$ and 'k' in (3) represents the dimensionality of the feature space.

Cosine Similarity represented as,

$$\cos\theta = \frac{\overrightarrow{C_1}.\overrightarrow{C_2}}{\|C_1\|\|C_2\|} \tag{4}$$

In (4), the numerator represents the dot product of two vectors C1 and C2, and the denominator represents the product of Euclidean distance. Minkowski is a generalized similarity measuring formula of Manhattan and Euclidean distance measurement. It is represented in (5).

$$d_{minkowski(x,y)} = \left(\sum_{i=1}^{k}(|x_i - y_i|)^r\right)^1 \tag{5}$$

In (3), x and y represents two points and 'r' value is a real values and always lies between 1 and 2. The default value of 'r' is 1. When r is 1 it is called Manhattan distance function and if r=2 it is called Euclidean distance function. Manhattan distance between two points on a line is represents as shown in (6). It is a similarity measure of real vectors.

$$d(x, y) = |x_i - y_i| + |x_i - y_i| \tag{6}$$

where x and y in (4) represents x is any point on line and y is a testing point.

Hamming Distance is one of the easiest representations in similarity measurement. It returns 1 if there is no similarity identified and returns 0 otherwise. It is represented as shown in (7). x_ds and y_ds are two data tuples in data set. It is a measure of similarity between binary vectors.

$$dist(x_{ds}, y_{ds}) = def \begin{cases} 0 \ if \ x_{ds} = y\_ds \\ 1 \ otherwise \end{cases} \tag{7}$$

Need of hamming distance: In the case of classification of categorical variables hamming distance is suitable. Based on the type of input variable it is very easy to select a specific type of distance function. In most of the cases Euclidean distance function is used to find the similarity measure [24]. Euclidean function is suitable when the input variables are similar. If input variables are not same, Hamming function is preferable. In the present paper the hamming distance code is implemented in kNN. Are considered. Hence the similarity between data tuples in k-dimensional space is measured with help of Hamming distance function.

With the above discussions it can be conclude that, the linear regression is mostly used when the target output is come close with straight line. It is easy to represent regression when the system contains single predictor. Where the straight line is not necessarily become straight line in non-linear like linear regression. The demerits of linear regression can be overcome with LDA. In non-linear regression the prediction model depends on independent values. The Naïve Bayes classifier can be suggested when there is a strong evidence of independent features of attributes. NB classifier further best fit with training examples with supervised training. The baseline for development of NB classifier is for text recognition, and automatic opinion prediction in politics, sports, movies, and etc. The performance of NB classifier is more advanced by including SVM. In NB classifier evaluating of maximum likelihood consumes linear time unlike other types of classifiers. In most of the classifier types iterative approximations are executed for estimation. The main advantage of NB classifier is it considers only few hypotheses from hypothesis space. But, there is a chance of diminishing the accuracy. The NB classifier can alter the dependencies exists between attributes.

The accuracy scaled with different ranges with different algorithms. The SVM is selected for better accuracy and most of the cases it produced the results with 87% [25], [26]. Comparatively the LDA and LR models are able to produce results with diminishing accuracy of 67% and 83% respectively. The NB classifier is also able to produce good accuracy 77% results when compared with above models [2].

Unlike NB classifier the kNN classifier has no training period hence it is called lazy learner. In the present work kNN classifer is chosen as it learns from training dataset while forecasting real time predictions. This feature demands kNN algorithm where faster responses are required. kNN algorithm does not require any training data at initial stage, new datasets can be added at real time without having any impact on accuracy in prediction. In the present work chunks of movie review data sets are used.

In kNN algorithm the computational complexity is high when there is a large volume of dataset. This increases the difficulty of finding the distance in every dimension of data. Normalization and standardization is done on the dataset before kNN is applied to any dataset. These imperfections lead the algorithm poor performance in opinion mining. The performance can be further increase by optimization schemes such as ABC algorithms

3. IMPLEMENTATION

In the present work the analysis is done with Python which is a high level strong, robust and dynamic programming language, and provide flexible and extensive library. It supports various Machine learning algorithms to evaluate the performance of various classification techniques. Pre-processing is an important step involved in text mining. The standard steps involved in opinion mining concept are shown in Figure 1.

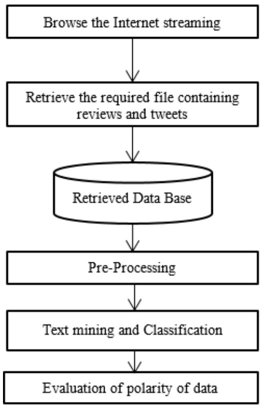

Browse the Internet streaming
Retrieve the required file containing reviews and tweets
Retrieved Data Base
Pre-Processing
Text mining and Classification
Evaluation of polarity of data

Figure 1. Data flow diagram of standard process for opinion mining

3.1. Levels of sentiment analysis

The level of sentiment analysis is very important and they are selected as per the user criteria. Various level of sentiment analysis is represented in Figure 2.

- Document level: Compute the polarity of individual sentence or word and combine all polarity values to append the polarity of the document.
- Sentence or phrase level: In this approach find the sentiment of individual word in the sentence and evaluate the polarity of the whole sentence.
- Feature level: In this approach the sentiment is extracted from the features of the product and by identifying the product.
- Word level: This is the mostly used approach, first one is dictionary based and second one is corpus based approach. In dictionary based approach the words are marked manually. The dictionary data set grows by searching synonyms and antonyms from dictionary. But this approach failed in dealing words with context specific or domain specific classification. In corpus based approach, words are added to dictionary with respective to domain specification. The set of related words grows in corpus dictionary by semantic or statistical techniques. Accuracy of such schemes can further optimized with parallel processing elements.

Levels of Sentiment Analysis

- ➢ Document level
- ➢ Sentence Level
- ➢ Word Level
- ➢ Feature Based

Figure 2. Levels of sentiment analysis

4. OPTIMIZATION TECHNIQUE FOR CLASSIFICATION

The classification techniques are applied on publicly available dataset like international movie review data base [4]. A small subset of IMDB is used for the purpose of evaluating the best classification technique in-order to analyse the various types of data. Different classification techniques are implemented here inorder to evaluate the best method which is predicting with more accurate results. In the present work it is observed that kNN is producing good results in-terms of accuracy. The accuracy obtained in kNN further improved with optimization technique artificial bee colony (ABC) algorithm. The average accuracy is measured by retrieving target snippets from the database contains user opinions on the movies is shown in Table 1. More than 100 datasets are used for training the system. The accuracy is calculated using the fundamental formulae shown in (8),

$$(T_p + T_n)/(T_p + F_p + T_n + F_n) \tag{8}$$

where, the T_p is true positive, T_n is true negative, F_p is false positive, and F_n is False negative. The true positive (T_p) is defined as document is identified as positive and recognised as positive. The False Positive (F_p) is defined as document is identified as positive but it is not positive. Similarly, the document is identified as negative and it is truly negative then they are called true negative (T_n). The document which is identified as negative but it is not really negative then they are called false negative (F_n).

In the present work unprocessed html files are collected, which contains 27886 user reviews on different movies downloaded in archive format. The IMDB data is broken into several datasets contains individual user review files. The whole database is decomposed into subsets of database in order to optimize the quality of the prediction [27]-[31]. Out of which 32 datasets are selected for testing. Individual dataset is tested using various linear and Non-linear algorithms.

In the present work mixed results in terms of accuracy are obtained with different algorithms implemented on different sizes of data sets. The average accuracy is calculated at each algorithm implemented on different data sets and is shown in Table 1. The accuracy and time complexity of KNN is further improved by testing with artificial bee colony (ABC) algorithm [32]-[34]. The time complexity increases as the k neighbours increases. The time complexity further optimized with continuous training. The ABC algorithm implementation procedure is explained in the following paragraphs. The results are presented in Figures 3 and 4.

Table 1. Average accuracy measured through various algorithms

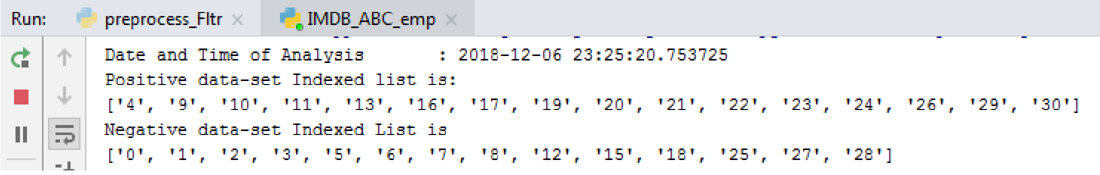
Name of Data Base	Name of the Algorithm	Average Accuracy (%)
IMDB	LR	67
	LDA	83
	NB	87
	SVM	91
	When k=3	
	KNN	76
	kNN with ABC	87
	When k=5	
	KNN	81
	kNN with ABC	87
	When k=20	
	KNN	86
	kNN with ABC	91
	When k=32	
	KNN	91
	kNN with ABC	94

4.1. ABC algorithm

The employee bees based on the selected attributes will carry the hotspots of the data that is the locations of the datasets to the hive. Based on the attributes the datasets are classified into categories like positive, negative, or neutral type data reviews. In Figure 3 the numbers are indicating the indexes of the identified dataset. The nearest neighbours with relevant features are clustered into one k-space.

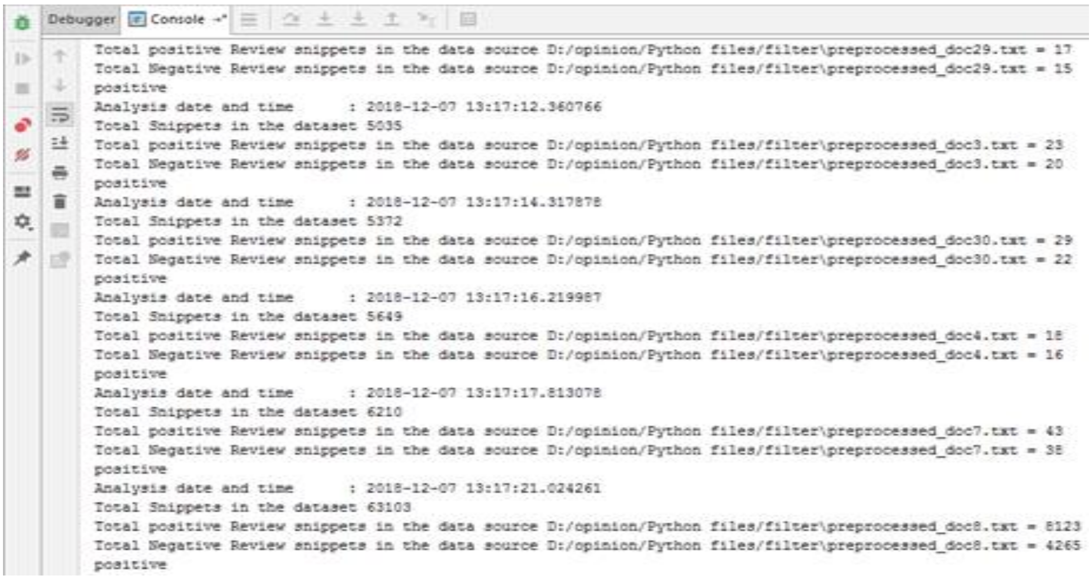
The employee bees memorize information and carries to the hive. Then onlookers in the hive will directly goes to the particular location in order to collect the required reviews from the data sets for analysis. It finally analysed the types of reviews whether they are positive, negative, or neutral. If the particular data set contains more positive words (positive dataset) then the result will be displayed as positive. If the particular file in the dataset contains more negative words then it display negative. The positive or negative oriented results are depends based on the attributes set at the classification. The results including the sum of all hypotheses are shown in Figure 4. The ABC can be implemented in association with KNN algorithm which can be combinable to enhance the accuracy in the predicted results.

For instance, when k is set to 20 then the ABC algorithm will collect all relevant and find the nearest neighbours based on its distance calculated with hamming distance. Figure 4 is showing a snap shot taken while the ABC algorithm is running to evaluate the positive reviews from the datasets. With proper training the employee will directly goes to the locations of the positive reviews. The onlooker results and employee memorization values can be matched from Figure 3 with Figure 4. For reader point of view the program is developed in such a way that the file name observed in Figure 4 is appended with the employee bee memorized value. For instance in the file name ‘preprocessed_doc29.txt’, the number 29 represents the employee collected (memorized) value from database. The ABC algorithm implemented for the present context is explored in Figure 5. The ABC algorithm is implemented on pre-processed data.



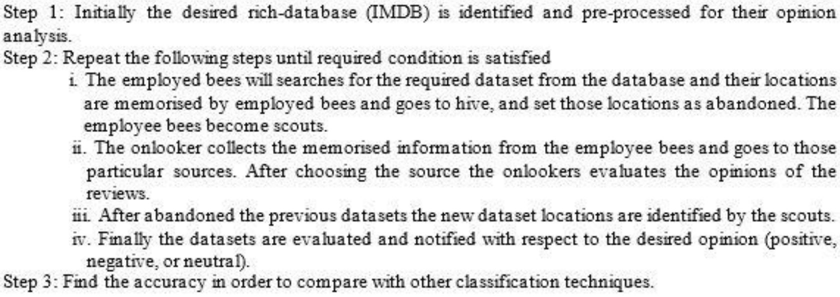
```
Date and Time of Analysis      : 2018-12-06 23:25:20.753725
Positive data-set Indexed list is:
['4', '9', '10', '11', '13', '16', '17', '19', '20', '21', '22', '23', '24', '26', '29', '30']
Negative data-set Indexed List is
['0', '1', '2', '3', '5', '6', '7', '8', '12', '15', '18', '25', '27', '28']
```

Figure 3. Hotspots of the data sets collected by employee bees



```
Total positive Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc29.txt = 17
Total Negative Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc29.txt = 15
positive
Analysis date and time      : 2018-12-07 13:17:12.360766
Total Snippets in the dataset 5035
Total positive Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc3.txt = 23
Total Negative Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc3.txt = 20
positive
Analysis date and time      : 2018-12-07 13:17:14.317878
Total Snippets in the dataset 5372
Total positive Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc30.txt = 29
Total Negative Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc30.txt = 22
positive
Analysis date and time      : 2018-12-07 13:17:16.219987
Total Snippets in the dataset 5649
Total positive Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc4.txt = 18
Total Negative Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc4.txt = 16
positive
Analysis date and time      : 2018-12-07 13:17:17.813078
Total Snippets in the dataset 6210
Total positive Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc7.txt = 43
Total Negative Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc7.txt = 38
positive
Analysis date and time      : 2018-12-07 13:17:21.024261
Total Snippets in the dataset 63103
Total positive Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc8.txt = 8123
Total Negative Review snippets in the data source D:/opinion/Python files/filter\preprocessed_doc8.txt = 4265
positive
```

Figure 4. Analysis of the opinion evaluated from the datasets



```
Step 1: Initially the desired rich-database (IMDB) is identified and pre-processed for their opinion analysis.
Step 2: Repeat the following steps until required condition is satisfied
    i. The employed bees will searches for the required dataset from the database and their locations
       are memorised by employed bees and goes to hive, and set those locations as abandoned. The
       employee bees become scouts.
    ii. The onlooker collects the memorised information from the employee bees and goes to those
        particular sources. After choosing the source the onlookers evaluates the opinions of the
        reviews.
    iii. After abandoned the previous datasets the new dataset locations are identified by the scouts.
    iv. Finally the datasets are evaluated and notified with respect to the desired opinion (positive,
        negative, or neutral).
Step 3: Find the accuracy in order to compare with other classification techniques.
```

Figure 5. ABC algorithm used for opinion mining

4.2. Random distance weighted ABC kNN algorithm

In the present work kNN algorithm is implemented with ABC non-parametric algorithm with hamming distance inorder to improve the accuracy. The method is proposed to optimize the weights for each data set of random query identified around nearest k neighbouring positions. The proposed algorithm divided into two stages. In the first stage the test or query data set is assigned with optimal weights (food source) at each k-value using ABC algorithm. In the second stage the kNN algorithm assigns the class to the query data. The proposed hamming distance based ABC kNN algorithm is shown below.

Step 1: Initialise colony size and k-value
Step 2: Assign index to each food source i.e for each data set.
Step 3: for each dataset to be test

- Compute the hamming distance to find the nearest neighbour (s)
- Run the ABC algorithm
- Memorize these values
- Store the best weights against the memorised values

Step 4: Assign the class label based on the best weights assigned by hamming distance code
Step 5: repeat step 3 and step 4 for all query sets
Step 6: end

5. CONCLUSION

In this paper the importance of classification of user reviews on movies is presented. In the present work the existing classification schemes such as LR, LDA, NB, SVM, kNN, distance based KNN-ABC are discussed and compared in terms of average accuracy. It is observed that out of all classification techniques kNN with ABC model is producing best accuracy readings. When comparing with other algorithms the kNN produced fair reading of accuracy 71% is obtained on IMDB datasets. Similarly the kNN classification with ABC is further implemented to analysis user opinion. kNN alone has accuracy of 76% classification and in the present work the results are optimised to 94% of accuracy with hamming distance based kNN in association with ABC. The optimization of searching rate and retrieving dataset which are best fit k-neighbours with the target function is achieved with ABC algorithm.

REFERENCES

[1] Y. N. Rao, N. S. Kumar, Ch Vytarani, Ch R. Babu, Gsvp Raju, "Mimicked Web Page Detection over Internet," *International Journal of Electronics Communication and Computer Engineering (IJECCE)*, vol. 5, no. 1, pp. 104-108, 2014.

[2] J. Padmavathi, L. Heena, F. Sabika, "Effectiveness of Support Vector Machines in Medical Data mining," *Journal of Communications Software and Systems (JCOMSS)*, vol. 11, no. 1, pp. 25-30, 2015, doi: 10.24138/jcomss.v11i1.114.

[3] S. Narra, L. N. Chaitanya, S. Ponugoti, N. S. Kumar, "Integrating and Organisation of Multidimensional Virtual Citizen Database with Extinction and Limited Access," *International Journal of Computer Applications (IJCA)*, vol. 85, no. 2, pp. 24-28, 2014, doi: 10.5120/14814-3035.

[4] "Movie Review Data," 2012. [Online]. Available: http://www.cs.cornell.edu/people/pabo/movie-review-data/polarity_html.zip.

[5] A. Scheidler & M. Middendorf, "Learning classifier systems to evolve classification rules for systems of memory constrained components," *Evolutionary Intelligence*, vol. 4, no. 3, pp 127-143, 2011.

[6] H. G. Ramaswamy, A. Tewari, S. Agarwal, "Consistent algorithms for multiclass classification with an abstain option," *Electronic Journal of Statistics*, vol. 12, no. 1, pp. 530-554, 2018, doi: 10.1214/17-EJS1388.

[7] P. Auer, A. Clark, T. Zeugmann, "Algorithmic Learning Theory," *Theoretical computer science*, vol. 650, pp. S.1-S.3, 2016, doi: 10.1016/j.tcs.2016.07.027.

[8] M. Schmidt, N. Le Roux and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, pp. 83-112, 2017, doi: 10.1007/s10107-016-1030-6.

[9] Vanashri *et al.*, "Logistic Regression: Aggregating Reviews by User Preference Modeling," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 3, no. 8, 2015.

[10] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2011, pp. 142-150, 2011.

[11] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol 26, no. 2, pp. 181-191, 2005, doi: 10.1016/j.patrec.2004.09.014.

[12] G. Sania, V. Suresh, "Opinion Mining on Twitter Data of Movie Reviews using R," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 19, no. 4, pp. 19-24, 2017, doi: 10.9790/0661-1904041924.

[13] A. O Adetunmbi, O. A. Sarumi, O. Boyinbode, “Machine Learning Approach to Sentiment Analysis of Users Movie Reviews,” *2nd International Conference on Information and Communication Technology and Its Applications (ICTA 2018)*, 2018, pp. 327-332.

[14] P. Baid, A. Gupta, N. Chaplot, “Sentiment Analysis of Movie Reviews using Machine Learning Techniques,” *International Journal of Computer Applications (IJCA)*, vol. 179, no.7, pp. 45-49, 2017, doi: 10.5120/ijca2017916005.

[15] H. Zang, “The optimality of Naïve-Bayes,” *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach, Florida, USA, pp. 1-6, 2004.

[16] P. Gamallo, M. Garcia, “Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets*,” 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 2014, pp. 171-175, doi: DOI:10.3115/v1/S14-2026.

[17] V. N. Vapnik, “The Natural of Statistical Learning Theory,” *Springer-Verlag New York*, 1995.

[18] V. N. Vapnik, “Statistical Learning Theory,” *Wiley, New York*, NY, USA, 1998.

[19] S. Liu, F. Li, F. Li, X.Cheng, & H. Shen, “Adaptive co-training SVM for sentiment classification on tweets,” *CIKM '13: Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2079-2088, doi: 10.1145/2505515.2505569.

[20] S. A. Dudani, “The distance weighted k-nearest-neighbor rule,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 4, pp. 325-327, 1976, doi: 10.1109/TSMC.1976.5408784.

[21] H. Yigit, “ABC-based distance-weighted kNN algorithm,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 27, no. 2, pp. 189-198, 2015, doi: 10.1080/0952813X.2014.924585.

[22] A. Moosavian, H. Ahmadi, A. Tabatabaeefar, M. Khazaee, “Comparison of two classifiers; K-nearest neighbour and artificial neural network, for fault diagnosis on a main engine journal-bearing,” *Shock and Vibration*, vol. 20, no. 2, pp. 263-272, 2013, doi: 10.1155/2013/360236.

[23] R. W. Hamming, "Error detecting and error correcting codes," *The Bell System Technical Journal*, vol. 29, no. 2, pp. 147-160, 1950, doi: 10.1002/j.1538-7305.1950.tb00463.x.

[24] B. Wang, X. Gan, X. Liu, B. Yu, R. Jia, L. Huang, & H. Jia, “A Novel Weighted KNN Algorithm Based on RSS Similarity and Position Distance for Wi-Fi Fingerprint Positioning,” *IEEE Access*, vol. 8, pp. 30591-30602, 2020, doi: 10.1109/ACCESS.2020.2973212.

[25] B. Pang, and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” *42nd Meeting of the Association for Computational Linguistics (ACL-04*), 2004, 271-278.

[26] A. K. Vishal, S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques,” *International Journal of Computer Applications (IJCA),* vol. 139, no. 11, pp. 5-15, 2016, doi: 10.5120/ijca2016908625.

[27] M. Abedini, M. Kirley, “Coxcs: a coevolutionary learning classifier based on feature space partitioning,” *AI 2009: Advances in Artificial Intelligence, 22nd Australasian Joint Conference,* Melbourne, Australia, vol. 5866, pp. 360-369, 2009, doi: 10.1007/978-3-642-10439-8_37.

[28] M. Gershoff, S.Schulenburg, “Collective behavior based hierarchical xcs,” *GECCO '07: Proceedings of the 9th annual conference companion on Genetic and evolutionary computation,* 2007, pp. 2695-2700, doi: 10.1145/1274000.1274064.

[29] U. Richter, H. Prothmann, H. Schmeck, “Improving XCS performance by distribution,” *SEAL '08: Proceedings of the 7th International Conference on Simulated Evolution and Learning*, vol. 5361, pp. 111-120, 2008, doi: 10.1007/978-3-540-89694-4_12.

[30] F. Zhu, S.-U. Guan, “Cooperative co-evolution of GA-based classifiers based on input decomposition,” *Engineering Applications of Artificial Intelligence*, vol. 21, no. 8, pp. 1360-1369, 2008, doi: 10.1016/j.engappai.2008.01.009.

[31] K. C., Santosh, “Document Processing using Machine Learning Techniques,” *CRC Press*, 2018.

[32] K. Dervis, “An idea based on honey bee swarm for numerical optimization,” *Techinical Report-TR06*, 2005.

[33] W. F. Gao, S. Y. Liu, L. L. Huang, “A Novel Artificial Bee Colony Algorithm Based on Modified Search Equation and Orthogonal Learning,” *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 1011-1024, 2013, doi: 10.1109/TSMCB.2012.2222373.

[34] Y. Wang, J. You, J. Hang, C. Li and L. Cheng, “An Improved Artificial Bee Colony (ABC) Algorithm with Advanced Search Abilityq,” *2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC),* Beijing, China, 2018, pp. 91-94, doi: 10.1109/ICEIEC.2018.8473513.