An Effective Hybrid Cuckoo Search with Harmony Search for Review Spam Detection

S.P.Rajamohana
Assistant Professor (Sr.Gr),
Department of IT,
PSG College of Technology,
Coimbatore,India.

K. Umamaheswari
Professor
Department of IT
PSG College of Technology,
Coimbatore,India.

S.Vasantha Keerthana
PG Scholar
Department of IT
PSG College of Technology
Coimbatore,India.

***Abstract* - In the recent years, online reviews are one of most important source of customer opinion. Nowadays consumer can gain knowledge about the products and service from online review resources, using which they can make decisions. This may lead to Opinion Spam, where spammers may manipulate and fake reviews to promote artificially or devalue the products and other services. Opinion spam detection is done by extracting meaningful features from the text, and identifying the spam reviews using machine learning techniques. This representation results in a very high dimensional feature space. These features are irrelevant, redundant, and noisy which may affect the performance of the classifier. Therefore, a good feature selection method is needed in order to speed up the processing rate, predictive accuracy. Evolutionary algorithms for feature selection can be used to handle these high-dimensional feature spaces which eliminate the noisy and irrelevant features. In this work, an effective hybrid feature selection technique using Cuckoo Search with Harmony search is proposed and Naive Bayes is used for classifying the review into spam and ham.**

Index Terms - Review Spam Detection, Feature Selection, Cuckoo search and Harmony search

I. INTRODUCTION

Nowadays everything has become very fast due to internet. As there are too many social networking sites hence people are interacting with each other across the world. They can share their ideas on internet. Also internet provides the facility of online shopping, so related to this on companies website or some review web sites such as Amazon, Ebay, flipkart, Yelp and many more provides lots of reviews about products. Before purchasing any product, it is a normal human behavior to do a survey on that. Hence these websites are helpful to the people to check quality of product. Based on available reviews customer can compare different brands of product and can buy a product. Hence these reviews will change the mind set of customer. If these reviews are true then it can help customer to select proper product satisfying their requirements. Similarly, if reviews are false or not true then it can yield wrong information to customers. The process of review spam detection involves using machine learning algorithms for spam detection. The performance of the algorithms can be improved by removing irrelevant, redundant and noisy features. Therefore a good feature selection method is used for improving the accuracy of machine learning algorithms .Temporal and spatial features are used for feature selection and supervised approach was proposed for opinion spam detection in [13] to perform opinion spam analysis using a large scale real time dataset. In [14], review graph method was applied to determine the relationships between reviewers, reviews. A hybrid PU-learning-based Spam Detection model was presented in [15] to detect multi-type spammers by means of adding or recognizing only a small portion of positive samples. In [16], to detect spam in Arabic reviews, new approach was designed which integrates data mining and text mining in classification approach. In [17], reviews spam detection was implemented for detecting reviews on brand spam detection. A novel approach was presented in [18] the spam reviewers are identified based on the writing and behavior styles of sentiment oriented text. Review spam detection method [19] was used to identify the store review spammer depends on review relationship. Chinese spam review detection system [20] was developed based on rules with aid of Naive Bayes Classifier for spam review detection. In [21], IPSO_NB based Feature selection was proposed for detecting fake reviews. Feature selection using Binary artificial colony with Naive Bayes was proposed in [22]. Shuffled frog leaping algorithm with Naive Bayes based Feature selection was developed for Sentiment Classification in [23]. Latent Dirichlet Allocation was applied for selecting features from the movie review dataset and SVM with SMO classifier was used for sentiment classification in [24]. Hybrid PSO_SVM approach was implemented for classifying the reviews into positive and negative [25].

Based on aforementioned methods and techniques presented, a novel Hybrid Cuckoo with Harmony search based Review Spam Detection method is proposed in this paper to efficiently detect the review spam detection.

II. METHODOLOGY

The proposed system is to identify the best features using hybrid cuckoo and harmony search and Naive Bayes classifier is used for classification. The below figure1 shows the overall flow of the proposed system.

978-1-5090-5434-3

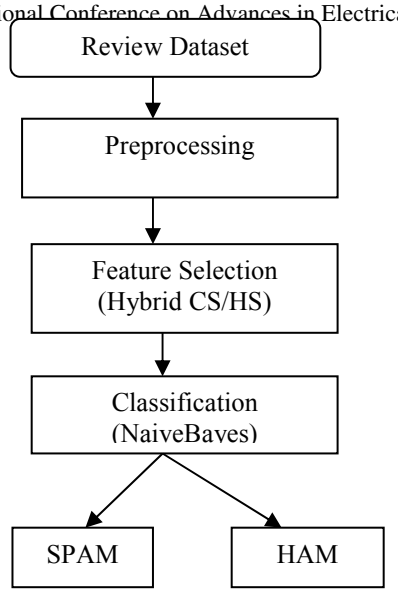

Figure 1 Methodology of Review spam detection

A. DATA PREPROCESSING

Initially, Data preprocessing is performed for changing the document into an appropriate format. The data preprocessing step contains the following processes such as tokenization, stemming and stop word removal. During the tokenization process, the texts are split into tokens such as words. In the process of tokenization, it removes question marks, punctuation and whitespaces. The most commonly used words in the dataset are removed with help of stop word removal process, because stop words are not useful for classification. Porter stemmer algorithm is used for removing the general morphological and inflexional endings from words in English. For example: classify, classifies and classifying. These three words taken from only one word "classify". Therefore, we consider only the root word "classify". Senitwordnet is used to extract the opinionated words. After determining the sentiment scores from SWN (i.e. features), we construct the TF-IDF Matrix for matching feature value in particular document. TF-IDF is the statistical technique which impersonates how vital a word is to be present in a document in a collection.

B. HYBRID FEATURE SELECTION USING CUCKOO WITH HARMONY SEARCH

Feature selection is applied to find a small number of relevant features to reach better classification accuracy than using the whole features. In this paper a hybrid feature selection using cuckoo with harmony search is used.

a.Cuckoo search

In hybrid cuckoo with harmony search is a new metaheuristic algorithm. In cuckoo search, each nest denotes a candidate solution to the problem and a set of nests are referred to as a population. Egg represents the features. Initially, population are generated randomly. The nests are represented as a binary string of 0s and 1s, where 1 represents the features are selected and else the other way if the features not selected.In each iteration, using random walk via Levy flights the nests are updated.

b. Harmony search

Harmony search is a new metaheuristic algorithm was proposed by Geem et al in 2001[1]. The parameter of the harmony search is the harmony memory, the harmony memory size, the harmony memory consideration rate, the pitch adjustment rate and the pitch adjustment bandwidth. The suitable representation of harmony memory is a two dimensional matrix .Here the rows represents harmonies (solution vectors) and the number of rows is predefined by the size of harmony memory. The columns represent the features generated from the review dataset. No of columns depends represent s the total number of features. Each column is dedicated to one musician; it not only stores the good notes previously played by the musician but also provides the pool of playable notes for future improvisations.

c. Hybrid Cuckoo with Harmony search

In general, the standard Cuckoo search algorithm explores the search space, and finds the global optimal value very quickly, but it exploits solutions poorly due to occasionally large steps. On the other hand, standard harmony search is well capable of exploiting solutions by carefully tuning PAR. In the proposed work, hybrid Cuckoo with harmony search (CS/HS) is proposed to find the optimized feature subset. In hybrid CS/HS method, the pitch adjustment rate(PAR)operator of Harmony search is introduced as a mutation operator in cuckoo search [3]. In this way, this method can explore the new search space by the hybrid HS operator of PAR and exploit the population with CS. In the exploitation stage, once an individual is chosen among the current best individuals, a new cuckoo individual is generated globally using Lévy flights. We fine tune every element using PAR operator of HS can exploit the advantages of both Cuckoo and Harmony search.

Algorithm 1: Hybrid Cuckoo with Harmony search

1. Input: Divide the dataset in to training data set and testing data set
2. begin
   Generate Initial population of host nest n with f eggs:
3. **while** max iterations *is not reached* **do**
4. Calculate the fitness for each nest fi using Naive Bayes Classifier
5. Update each nest using cuckoo levy flight,
6. $xji\ (t) = xji\ (t - 1) + \alpha \oplus Levy(\lambda)$
7. In hybrid approach, PAR operator is applied for r mutating the updated solution.
8. For(j=1 to n)
9. $If(\epsilon 1 < HMCR)$ // $\epsilon 1$
10. $xji(t) = y * xji(t)$

978-1-5090-5434-3©2017 IEEE

11. $If(\epsilon 2 < PAR)$ // $\epsilon 2$
12. $xji(t) = xji(t) + bw * (2 * rand - 1)$
13. End if
14. Else
15. Keep the existing solution
16. End if
17. End for
18. Eliminate the worst nest with probability P_a
19. Pass the best nests to next iteration
20. End

d .Fitness Function

Classification accuracy plays a vital role in the process of spam review classification using significant features from the feature vector. The performance of any classification system is measured by its classification accuracy. Here, accuracy of the kNN classifier is used as the fitness function.The fitness function of hybrid cuckoo and harmony search uses the following formula.

Fitness = accuracy () (1)

III.RESULS AND ANALYSIS

The dataset used in the experiment is spam review dataset. The dataset consists of 1600 labeled examples of deceptive and truthful review about the 20 Chicago hotels. The corpora consists of 800 truthful reviews, 800 deceptive reviews. Hotel review dataset have been considered for spam review classification. The proposed system has been implemented using java with weka. Features are selected using hybrid cuckoo with harmony search. The classification accuracy of kNN is used as the fitness function of CS/HS. The Cuckoo search, the number of nest is 30 and it was run for 500 iterations. The value of *pa*, α and λ are set as 0.3, 1 and 1.5 (Gunavathi et al., 2015),(Gai-Ge Wang et al.,2016). The parameters used for the hybrid cuckoo and harmony search is shown in the Table 1.

Table 1 Parameter Settings for Hybrid Cuckoo with Harmony Search

Parameters	Values
No of nests	30
pa	0.3
α	0.1
λ	1.5
HMCR	0.9
PAR	0.1

In the proposed work, initially 1750 features were extracted using TFIDF. Preprocessing results are shown in Table 5. Feature Selection is carried out using hybrid cuckoo and harmony search to obtain the reduced feature subset. The results are compared with binary cuckoo search results. The Table 6 shows that the number of features obtained by CS/HS is lesser than the number of features obtained by binary CS.

Table 2 Preprocessing Results

Total no of Reviews	1600
No of features	589056
After Stop word removal	476140
After Stemming	242748
Sentiwordnet	129838
Duplicate feature removal	110090
TFIDF	1750

Table 3 Feature selection Results

Algorithm	Total number of features
Binary Cuckoo Search	1355
Hybrid Cuckoo and Harmony search	1015

The proposed cuckoo with harmony search selects totally 1015 features and provides 91.12 % average classifier accuracy by using Naive Bayes and selects 1355 features and provides 82.34 % average classifier accuracy by using kNN classifier. The comparison results of classification accuracy using kNN and Naive Bayesare shown in Table 4.

Table 4 Performance Analysis of Classification Accuracy using kNN and Naive Bayes

Review	KNN	Naive Bayes
400	77.11	87.21
800	79.99	89.67
1200	80.91	90.23
1600	82.34	91.12

Table 5 shows that the global fitness values obtained for hybrid approach is higher when compared to binary cuckoo search and also the fitness value increases as the number of iteration increases.

Table 5 Fitness Comparison of Binary Cuckoo Search and Hybrid Approach

No of Iterations	Global best fitness	Global best fitness
	Binary Cuckoo Search	Hybrid approach
100	0.86922	0.89998
200	0.87794	0.91091
300	0.90347	0.92467
400	0.91113	0.93990
500	0.92100	0.94190

Table 6 Results of Classification

Classifier	SPAM	HAM
Naive Bayes	620	980
kNN	680	920

978-1-5090-5434-3©2017 IEEE

Table 6 shows that out of 1600 reviews taken the Naive Bayes classifer have identified 620 reviews as spam reviews and 980 as ham reviews. The kNN classifier has identified 680 reviews as spam reviews and 920 as ham reviews.

IV. CONCLUSION

In this work, we have proposed a hybrid cuckoo with harmony search for feature selection, to select the optimized feature subset from the dataset and Naive Bayes is used for classification. Experimental results show that the hybrid cuckoo with harmony search is capable of identifying good quality feature subsets. The resulting classification accuracy tested and is compared with the whole feature subset. In almost all aspects, the proposed approach delivered considerably better results than binary cuckoo search. The proposed hybrid feature selection method for review spam detection provides better classification accuracy with an optimized feature subset.

REFERENCES

[1] Ren Diao and Qiang Shen, "Feature Selection with Harmony Search", IEEE Transactions on Systems, Cybernetics, December 2012.

[2] D. Rodrigues, L. A. M. Pereira, T. N. S. Almeida, J. P. Papa, "Binary Cuckoo Search Algorithm for Feature Selection" IEEE International Conference on Systems, Man, and Cybernetics, 2014.

[3] Gai-Ge Wang, Amir H. Gandomi ,Xiangjun Zhao "Hybridizing Harmony Search Algorithm with Cuckoo Search For Global Numerical Optimization", Springer-Verlag Berlin Heidelberg, 2014.

[4] Patchara Nasa, Khamron sunat, sirrpat Chewchanwattana "Enhancing Modified Cuckoo Search using Mantegna Levy Flight and Chaotic Sequence" International Joint Conference on Computer Science and Software Engineering, 2013.

[5] Ahmed Abbasi, Stephen France, Zhu Zhang, and Hsinchun Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks", IEEE Transactions on Knowledge and Data Engineering, 2011.

[6] Bandakkanavar RV, Ramesh M, Geeta H "A Survey On Detection Of Reviews Using Sentiment Classification of Methods", 2014.

[7] Abbasi A, Zhang Z, Zimbra D, Chen H, Nunamaker JF "Detecting Fake Websites: The Contribution of Statistical Learning Theory", 2010.

[8] Jindal N, Liu B, Lim EP "Finding Unusual Review Patterns Using Unexpected Rules". In: Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[9] Li F, Huang M, Yang Y, Zhu X "Learning To Identify Review Spam", IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 2011.

[10] Mukherjee A, Liu B, Glance N "Spotting Fake Reviewer Groups In Consumer Reviews". In: Proceedings of the 21st International Conference on World Wide Web, ACM, 2012.

[11] Shojaee S, Murad MAA, Bin Azman A, Sharef NM, Nadali S "Detecting Deceptive Reviews using Lexical and Syntactic Features". In: Intelligent Systems Design and Applications (ISDA), IEEE, 2013.

[12] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, 1997.

[13] Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, 2003.

[14] Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu, "Identify Online Store Review Spammers via Social Review Graph", ACM Transactions on Intelligent Systems and Technology, 2012.

[15] Zhiang Wu, Youquan Wang, Yaqiong Wang, Junjie Wu, Jie Cao1, Lu Zhang, "Spammers Detection from Product Reviews: A Hybrid Model", IEEE International Conference on Data Mining (ICDM), 2015.

[16] Ahmed Abu Hammad, Alaa El-Halees, "An Approach for Detecting Spam in Arabic Opinion Reviews", The International Arab Journal of Information Technology, 2015

[17] M.S.Patil, A.M.Bagade, "Review on Brand Spam Detection Using Feature Selection", International Journal of Advanced Research in Computer Science and Software Engineering", 2013

[18] Junlong Huang, Tieyun Qian, Guoliang He, Ming Zhong, Qingxi Peng, "Detecting Professional Spam Reviewers", Advanced Data Mining and Applications, Springer, 2013.

[19] Qingxi Peng, "Store Review Spammer Detection Based on Review Relationship", Advances in Conceptual Modeling, Springer, 2014.

[20] Xiujuan Xu, Tianqi Han, Zhenlong Xu, Yu Wang, Yu Liu, "Design and Implementation of Chinese Spam Review Detection System", Springer, 2013.

[21] SP.Rajamohana and Dr.K.Umamaheswari, "An Integrated Evolutionary Algorithm for Review Spam Detection on Online Reviews", Advances in Natural Applied Science (AENSI), 2016.

[22] SP.Rajamohana, Dr.K.Umamaheswari, " Feature selection using binary artificial bee colony for sentiment classification", International Research Journal of Engineering and Technology, 2016.

[23] SP.Rajamohana, Dr.K.Umamaheswari, "Sentiment Classification using Shuffled Frog Leaping Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, 2016.

[24] SP.Rajamohana and K.Umamaheswari, "Sentiment classification based on LDA using SMO classifier", International Journal of Applied Engineering Research, 2015.

[25] K.Umamaheswari, SP.Rajamohana, "Opinion mining using hybrid methods", International Journal of Computer Application, 2015.

978-1-5090-5434-3