Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 9, No. 2, May 2018), Pages: 37-48
www.jacr.iausari.ac.ir

An Improved K-Nearest Neighbor with Crow Search Algorithm for Feature Selection in Text Documents Classification

Ali Allahverdipoor, Farhad Soleimanian Gharehchopogh✉
Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
aliam1364@yahoo.com; bonab.farhad@gmail.com

Received: 2016/01/13; Accepted: 2016/06/07

Abstract

The Internet provides easy access to a kind of library resources. However, classification of documents from a large amount of data is still an issue and demands time and energy to find certain documents. Classification of similar documents in specific classes of data can reduce the time for searching the required data, particularly text documents. This is further facilitated by using Artificial Intelligence (AI) and optimization algorithms which are highly potential in Feature Selection (FS) and words extraction. In this paper Crow Search Algorithm (CSA) is used for FS and K-Nearest Neighbor (KNN) for classification. Additionally, TF technique is proposed for counting words and calculating the words' frequency. Analysis is performed on Reuters-21578, Webkb and Cade 12 datasets. The results indicate that the proposed model is more accurate in classification than KNN model and, show greater F-Measure compared to KNN and C4.5. Moreover, by using FS, the proposed model promotes classification accuracy by %27, compared to KNN.

Keywords: *Text Documents Classification, Crow Search Algorithm, K-Nearest Neighbor*

1. Introduction

Before of expansion of the internet and computer networks, the major challenge for people was lack of access to and evaluation of data [1, 2]. Today, nevertheless, the main problem is to find proper information from a bulk of available data. The widespread dominance of web and increasing amount of information call for methods and techniques that facilitate data processing. Text-mining and text processing methods are suggested in response to such overwhelming problems [3, 4]. Classification of text documents is a method for text processing whereby similar texts are categorized on the basis of title and text body [5]. Machine learning methods with high accuracy and detection are usually used in text classification and processing [6, 7].

Text documents contain information about a variety of topics in the form of sentences and paragraphs. They also contain specific words that help detect text type. Nevertheless, classification and detection of a large number of text documents are possible only by intelligent systems [8]. To detect a text, it is converted to a vector and then weighted [9]. Extraction of key words is essential in analyzing text documents and determines overall efficiency of text analysis [10]. Any lapses in extraction phase leads to further detection errors. Keywords are mapped into a two-dimensional feature space and then weighted to decrease the likelihood of classification errors [11].

Classification algorithms need to be accurate and high-quality in terms of validation. Meta-heuristic algorithms are highly potent in finding optimal solutions. They use information and experience gained during the search process as the memory for guiding the process of more prominent solution spaces. Thus, the present paper uses a combination of CSA [12] and KNN [13] to classify text documents. CSA is a meta-heuristic algorithm adopted from the life of crows for FS. Crows look for food in groups and are highly intelligent. KNN is a data-mining procedure that is used for classification of text documents. It proves to be powerful in detecting classes of data based on k-value [14, 15, 16 and 17].

The paper is structured as follows: in Section 2, we review the literature of the studies previously done on text documents classification. In Section 3, explains CSA, KNN and the proposed model. In Section 4, evaluates results and compares the proposed model with other models. Finally, in Section 5, conclusions and suggestions are made for future studies.

2. Related Works

There is a bulk of literature on text documents classification and FS, using different methods for FS and weighting. Some of these studies are given below.

Support Vector Machine (SVM) and Naïve Bayes (NB) are data mining algorithms proposed for classification of text documents [18] based on Improved Global Feature Selection Scheme (IGFSS) model that is a combination of FS models. Information Gain (IG), Gini Index (GI), Distinguishing Feature Selector (DFS), and Odds Ratio (OR) are used for FS in NB and SVM. Evaluations are done on Reuters-21578, Webkb and Classic3.

$$IG(t) = -\sum_{i=1}^{M} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{M} P(C_i \mid t)\log P(C_i \mid t) + P(\bar{t})\sum_{i=1}^{M} P(C_i \mid \bar{t})\log P(C_i \mid \bar{t}) \quad (1)$$

$$GI(t) = \sum_{i=1}^{M} P(t \mid C_i)^2 P(C_i \mid t)^2 \quad (2)$$

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i \mid t)}{P(\bar{t} \mid C_i) + P(t \mid \overline{C}_i) + 1} \quad (3)$$

$$OR(t, C_i) = \log \frac{P(t \mid C_i)[1 - P(t \mid \overline{C}_i)]}{[1 - P(t \mid C_i)]P(t \mid \overline{C}_i)} \quad (4)$$

Results show that combined models of IG+IGFSS, GI+IGFSS, and DFS+IGFSS are more accurate for classification than IG, GI, and DFS.

Text Classification using Semi-supervised Clustering (TESC) [19] is proposed, based on NB, for Reuters-21578 and TanCorp. It uses class centrality to detect key words and correlation between documents in order to promote accuracy. First, documents are randomly classified and a document is selected as a paradigm for selecting other documents. Results demonstrate that TESC is more accurate than Back-Propagation Neural Network (BPNN) and SVM.

B-Tree [20] is proposed for classification of text documents, compared to SVM, KNN and NB. B-tree is in fact a tree with roots at the top and leaves at the bottom level. Similar documents are weighted with leaves. B-Tree shows an accuracy of 87.85 on 20 Newsgroup, as compared to other models like SVM (85.65), KNN (70), and NB

(86.50). The obtained Accuracy on Google Newsgroup for B-Tree, SVM, KNN, and NB was 96.00, 45.25, 46.25, and 80.00, respectively.

Different techniques for words frequency have been proposed for classification of text documents on 20 Newsgroup, Reuters-21578, and TanCorp based on SVM and KNN [21]. Frequency is estimated using TF, TF-IDF, TF-IDF-ICDDF, TF-CHI, TF-Prob, TF-RF, TF-IGM, and RTF-IGM. TF-IGM and RTF-IGM show greater Accuracy and are defined by Equations (5) and (6).

$$w(t_k, d) = tf_{kd}.\left(1 + \lambda.\frac{f_{k1}}{\sum_{r=1}^{m} f_{kr}.r}\right) \quad (5)$$

$$w(t_k, d) = \sqrt{tf_{kd}}.\left(1 + \lambda.\frac{f_{k1}}{\sum_{r=1}^{m} f_{kr}.r}\right) \quad (6)$$

Results show that RTF-IMG in SVN and KNN shows greater accuracy in finding frequency than other methods. However, maximum accuracy was found for RTF-IGM in TanCorp. TF-IGM ranked second in terms of Accuracy.

Text Categorization using Feature Projections (TCFP), including SVM, NB, and KNN, is proposed for classification of text documents [22]. It uses synthetic methods to find frequency of key words. Evaluations are performed on Newsgroups, Reuters, and Webkb. F-measure for all the three data sets in TCFP is 86.19, 75.47, and 89.09, respectively, which is higher than SVM, NB, and KNN. Moreover, F-Measure for SVM in the above data sets was 82.49, 73.74, and 87.41, respectively.

Self-Organizing Maps (SOM) and SVM model is proposed for classification of text documents [23]. SOM is an unsupervised method based on competitive learning where units compete with each other in every single learning phase to remain active. Finally, the winning unit is weighted different from other units. Evaluations are performed on Reuters-21578, Webkb, 20 Newsgroup. The Results show that accuracy of SOM for these data sets was 60.14, 92.28, and 97.00, respectively, while the Accuracy of SVM was found to be 96.68, 94.73, and 99.72.

X^2, IG, MI, ECE, and t-test are methods used for extracting keywords [24]. SVM and KNN are also used for classification of Reuters-21578 and 20 Newsgroup. In t-test, data are characterized by relative or interval scale. Results reveal that SVM in X^2 method is more accurate than KNN. MI on 20 Newsgroup shows greater Accuracy. Finally, t-test is more efficient than other methods for vast documents.

KNN and C4.5, combined by Genetic Algorithm (GA), are proposed for classification of text documents in conjunction with frequency methods [25]. GA is used for FS. Evaluations are performed on Reuters-21578 and Classic3. Results of 7542 features demonstrate that F-Measure for KNN and C4.5 is 83.02 and 86.88, indicating that C4.5 is more accurate than KNN.

Different classification models are compared in Table (1).

Table 1. Comparison of Models for Classifying Text Documents

References	Models	Data sets	Technique	FS	Correlation	Time Complexity
[18]	SVM NB	●Reuters-21578 ●Webkb ●Classic3	Supervised	X	High	Low
[19]	NB	●Reuters-21578	Supervised	X	Medium	Low
[20]	B-Tree	●20 Newsgroup ●Google Newsgroup	Unsupervised	X	Medium	High
[21]	SVM KNN	●20 Newsgroup ●Reuters-21578 ●TanCorp	Supervised	X	High	Low
[22]	SVM NB KNN	●20 Newsgroup ●Reuters-21578 ●Webkb	Supervised	X	High	Low
[23]	SVM SOM	●20 Newsgroup ●Reuters-21578 ●Webkb	Supervised	X √	Medium	High
[24]	SVM KNN	●20 Newsgroup ●Reuters-21578	Supervised	X	High	Low
[25]	KNN C4.5	●Reuters-21578 ●Classic3	Supervised	√	High	Low

3. The Proposed Model

Efficiency, accuracy and speed are key factors for measuring similarity in text documents [26]. Detecting words frequency is highly important for extracting prominent features and accuracy of detections. TF is the most efficient scheme for weighing words in vector space of documents. One of the main problems that exist in the FS is that most of linear models are unable to recognize the best features and they cannot choose the graceful features. Therefore, selecting an appropriate algorithm for evaluating is an important criterion in discovering the best solution for the best category. The present paper proposes a synthetic model of CSA and KNN for classification of text documents. CSA is a meta-heuristic algorithm adopted from collective life of crows. In CSA, initial population is defined as $x^{i,iter+1} = [x_1^{i,iter}, x_2^{i,iter}, ..., x_d^{i,iter}]$. The position of each crow is defined by Equation (7) [12]. The initial population is carried out based on the amount of weight found by counting keywords. This means that at first, the total number of keywords are identified and then they take on a weight based on their frequencies. Vectors are distinguished based on weight and the vectors that have much more weight are elected as optimal vector.

$$Crows = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_d^1 \\ x_1^2 & x_2^2 & \dots & x_d^2 \\ . & . & . & . \\ x_1^N & x_2^N & \dots & x_d^N \end{bmatrix} \quad (7)$$

A number of N crows are randomly distributed in a d-dimensional space, each searching their surroundings. Crow memories are predefined and can improve their memories through experience and finding optimal solutions. Crow memory is defined by Equation (8) [12].

$$Memory = \begin{bmatrix} m_1^1 & m_2^1 & \dots & m_d^1 \\ m_1^2 & m_2^2 & \dots & m_d^2 \\ . & . & . & . \\ m_1^N & m_2^N & \dots & m_d^N \end{bmatrix} \quad (8)$$

The memories are updated by Equation (9) [12], where x is position of the crow, r is a random number [0, 1], f is flight distance and m is finding a new location.

$$x^{i,iter+1} = x^{i,iter} + r_i \times f^{i,iter} \times (m^{j,iter} - x^{i,iter}) \quad (9)$$

In the proposed model, data are summoned from Reuters-21578 [27], Webkb [28], and Cade 12 [28] and enter pre-processing filter where irrelevant and redundant words are removed to obtain higher accuracy. In fact, preprocessing is performed to purify documents off saturated data. Then, words extraction is performed using TF which prevents words scattering and is a powerful counter. Equation (10) is used for counting the words where (t_k, d_i) indicates frequency of each feature t_k in document d_i [29].

$$w_{ki} = tf(t_k, d_i) = \begin{cases} (t_k, d_i) & t_k \in vector\ of\ d_i \\ 0 & t_k \notin vector\ of\ d_i \end{cases} \quad (10)$$

Next, words are weighted to be analyzed smoothly in the features vector and to increase accuracy of detecting nearest neighborhood. When preprocessing, extraction and weighting are done, FS starts. Figure (1) shows the flowchart for the proposed model.

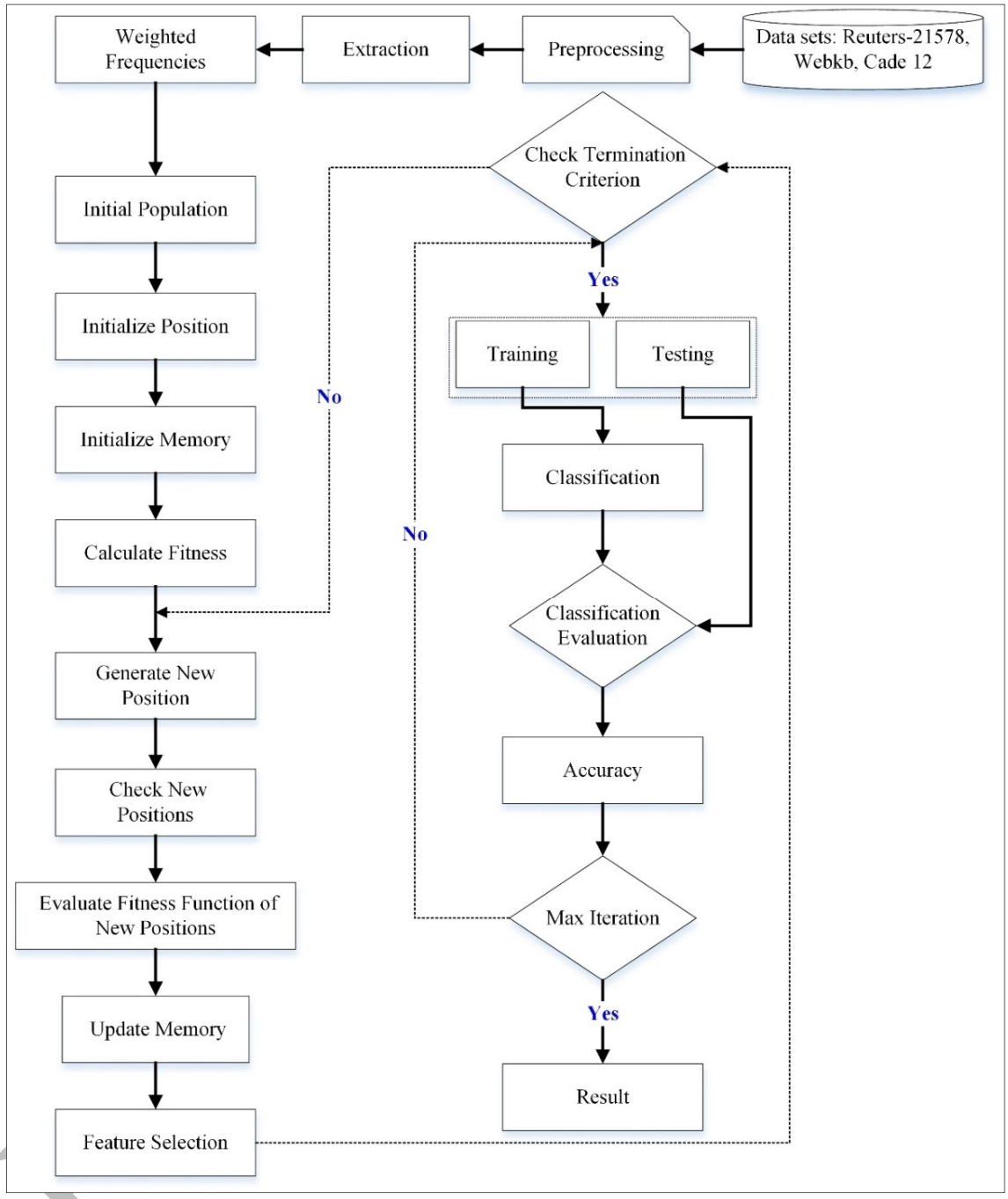

Figure 1. Flowchart for the Proposed Model

CSA is used for selecting optimal features to perform a better classification. Vector lengths are filled with words weights. First, initial population, which is weight of words, is placed on the number of vectors to obtain vector fitness. A vector with optimal fitness is selected and directed to KNN classifier, where the neighborhood of words weights is determined and the documents are trained by words weights. An unclassified sample can be easily identified by comparing it to other similar samples of KNN. Therefore, a measure is required to denote the distance between the samples. Suppose a feature vector $< a_1(x), a_2(x), \ldots a_n(x) >$. We use Euclidean distance of Equation (11) to determine the distance between x_i and x_j [13].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \quad (11)$$

Neighborhood of weights is estimated by Equation (11), and main key word is retrieved by its index, and the related documents are classified in the same cluster.

4. Evaluation and Results

The proposed model is evaluated on Reuters-21578, Webkb, and Cade 12 using VC#.NET 2015. Reuters-21578 documents contain 21578 texts in 135 different classes. Webkb documents includes a set of different documents on the internet, collected from different Computer Science Departments of universities, and contains 4199 text documents. Cade 12 documents include 40983 texts collected from web pages in Brazil. Table (2) represents results of KNN model on Reuters-21578, Webkb, and Cade 12 with different k values. When training is over, test documents are evaluated and detection accuracy is estimated using Equation (15) [30].

$$\Pr ecision = \frac{TP}{TP + FP} \quad (12)$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \quad (13)$$

$$F - Measure = \frac{2 * \Pr ecision * \mathrm{Re}\, call}{(\Pr ecision + \mathrm{Re}\, call)} \quad (14)$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (15)$$

TN indicates number of records whose real class is negative and the classifier algorithm recognizes them as negative. TP indicates number of records whose real class is positive and the classifier algorithm recognizes them as positive. FP indicates number of records whose real class is negative but the classifier algorithm recognizes them as positive. FN indicates number of records whose real class is positive but the classifier algorithm recognizes them as negative.

Table 2. Results of KNN model with different k values

Criterions	K	Data sets		
		Reuters-21578	Webkb	Cade 12
Precision	3	76.34	77.35	75.48
	4	73.24	73.87	72.00
	5	70.08	69.34	68.46
Recall	3	69.47	68.24	69.58
	4	67.28	65.07	67.32
	5	65.12	62.01	66.41
F-Measure	3	72.74	72.51	72.41
	4	70.13	69.19	69.58
	5	67.51	65.47	67.42
Accuracy	3	68.32	70.64	72.23
	4	65.02	67.32	70.31
	5	62.45	62.20	69.16

As can be seen in Table (2), k value is essential for detection accuracy. Accuracy is higher for the three datasets when k=3. However, maximum accuracy is found for Cade 12 (72.23). Table (3) shows results of the proposed model on Reuters-21578, Webkb, and Cade 12 with k and FS numbers.

Table 3. Results of the Proposed Model with k and FS Numbers

Criterions	K	FS=80			FS=120		
		Reuters-21578	Webkb	Cade 12	Reuters-21578	Webkb	Cade 12
Precision	3	94.34	96.47	95.30	95.30	96.34	97.34
	4	93.00	95.24	93.07	96.00	94.15	94.90
	5	91.27	93.03	91.56	95.18	92.27	93.27
Recall	3	79.32	76.34	80.14	82.02	86.30	85.47
	4	76.24	74.21	79.33	81.15	82.24	84.19
	5	74.20	73.41	76.84	80.20	80.20	81.22
F-Measure	3	86.16	85.23	87.07	88.16	91.04	91.02
	4	83.79	83.42	85.65	87.95	87.79	89.22
	5	81.85	82.06	83.56	87.05	85.81	86.83
Accuracy	3	96.24	97.54	96.34	97.24	96.00	96.86
	4	93.62	96.20	94.11	95.62	95.12	94.25
	5	91.84	94.06	92.77	92.84	93.17	93.64

Table (3) shows that F-Measure and Accuracy are more efficient for FS=120 because of increased feature diversity. Here, features are compared with more neighbors and common features in the same document are more likely to be found.

Table (4) compares the proposed model with KNN and C4.5 in terms of applied methods and number of features for FS on Reuters-21578 and k=3.

Table 4. Comparison of the Proposed Model with KNN and C4.5 on Reuters-21578

FS=IG [25]	KNN [25]			C4.5 [25]			Proposed Model		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
75	95.14	94.71	94.93	94.50	94.63	94.57	96.33	95.20	**95.76**
151	94.26	97.38	95.80	94.82	94.84	94.83	95.23	96.47	**95.85**
226	94.03	97.62	95.79	94.86	94.63	94.74	97.21	96.55	**96.88**
302	94.87	97.86	96.34	95.48	95.38	95.43	96.04	97.12	**96.58**
377	94.04	97.73	95.85	94.83	96.02	95.42	95.54	96.89	**96.21**
453	93.33	97.54	95.39	95.61	95.40	95.51	92.87	94.57	93.71
528	91.74	97.86	94.70	95.21	94.47	94.84	91.54	93.35	92.44
603	91.07	97.78	94.31	95.32	95.27	95.30	91.87	94.48	93.16
679	90.63	97.92	94.13	95.24	94.52	94.88	91.00	93.84	92.40
754	90.14	97.65	93.74	95.18	95.43	95.30	93.84	96.80	**95.34**
FS=IG -GA [25]	KNN [25]			C4.5 [25]			Proposed Model		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
42	95.37	94.68	95.03	96.20	93.40	94.78	94.23	93.12	93.67
83	96.64	95.99	96.31	95.98	94.47	95.22	96.94	95.88	**96.41**
121	97.50	96.93	97.21	95.39	94.60	95.00	97.68	95.47	96.56
169	98.17	97.52	97.84	95.95	95.65	95.80	97.82	98.90	**98.35**
197	97.73	97.60	97.66	96.41	95.40	95.90	96.54	96.00	96.27
241	97.42	97.73	97.57	96.51	95.40	95.96	98.84	96.67	**97.74**
286	97.16	97.84	97.50	96.11	95.65	95.88	95.24	94.27	94.57
317	97.04	98.05	97.54	96.40	95.11	95.75	96.03	95.66	95.84
352	97.04	98.10	97.57	95.84	96.08	95.96	99.00	96.47	**97.72**
384	96.93	97.78	97.35	95.72	95.51	95.61	97.40	96.50	96.95
FS=IG -PCA [25]	KNN [25]			C4.5 [25]			Proposed Model		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
36	93.97	93.75	93.86	95.68	93.53	94.60	96.35	95.44	**95.89**
71	96.47	95.54	96.00	95.17	94.76	94.97	97.35	96.24	**96.79**
103	96.91	96.37	96.64	94.96	94.66	94.81	96.20	96.00	96.10
134	97.77	97.03	97.40	95.66	95.40	95.53	95.24	94.78	95.01
162	97.24	97.11	97.18	95.54	95.62	95.58	98.21	97.68	**97.94**
193	96.87	97.46	97.16	95.70	95.67	95.68	97.63	98.11	**97.86**
222	94.70	94.70	97.16	95.68	95.35	95.52	95.37	94.89	95.13
250	96.80	97.73	97.26	95.76	95.32	95.54	98.02	97.31	**97.66**
278	96.29	97.76	97.02	95.66	95.40	95.53	96.80	98.54	**96.66**
303	96.34	97.65	96.99	95.48	95.35	95.42	94.10	93.54	93.82

It is clearly seen in Table (4) that Accuracy of the proposed model is greater than KNN and C4.5 when FS= IG for features of 75, 151, 226, and 302. In case of FS=IG-GA for features of 83, 241, 352, greater Accuracy is obtained. When FS=IG=PCA, the proposed model shows greater Accuracy for features of 36, 71, 162, 250. Moreover, KNN shows greater detection Accuracy than C4.5. In the proposed model, FS has been used to increase the accuracy and proximity of data, then the TP between the data is high. But in KNN and C4.5 models, the total space and the features of problem have been used, then the probability of FP is high among the categories. The reason why some features have much worse value is that CSA is stuck in some cases in the optimum local and there is no chance of reaching to the graceful precision. Also,

because some feature of vectors may not have adequate weight and therefore category model cannot very well find a similarity between categories.

The main advantages of CSA compared to other models is the use of memory in a way the graceful features are preserved by all crows. In other words, the CSA, each crows benefited from his past information, while there is no such behavior or trait in other algorithms, for example, in GA, there is no such memory and the prior knowledge of the issue is once lost with the population change. In CSA, each crow changes his position due to his personal experiences and the experiences of all the crows. As a result, there is fruitful cooperation between the crows and the crows in groups share their information with each other and find their best position that is the same weight.

5. Conclusion and Future Works

This paper proposed a synthetic model based on KNN and CSA. Evaluation of the model is performed on Reuters-21578, Webkb, and Cade 12. Our results indicate that the proposed model is more accurate than KNN it applies FS which promotes efficiency of classification and contributes too identifying weights of keywords in neighboring documents using KNN. The proposed model is also compared to KNN and C4.5 and improves F-Measure to %3. The paper seeks to promote accuracy of classification using FS based on CSA. However, optimal features selection is the major drawback in classifying text documents. We hope to come up with more optimal solution to FS on the basis of machine learning techniques.

References

[1] M. Rostami, S.S. Ayat, I. Attarzadeh, F. Saghari, Proposing a Method to Classify Texts Using Data Mining, Journal of Advances in Computer Research, Vol. 6, Issue 4, pp. 125-137, 2015.

[2] A.H. Jadidinejad, V. Marza, Building Semantic Kernel for Persian Text Classification with a Small Amount of Training Data, Journal of Advances in Computer Research, Vol. 6, Issue 1, pp. 125-136, 2015.

[3] A. Allahverdipour, F.S. Gharehchopogh, A New Hybrid Model of K-Means and Naïve Bayes Algorithms for Feature Selection in Text Documents Categorization, Journal of Advances in Computer Research, Vol: 8, No: 4, 2017(in press).

[4] K.N. Junejo, A. Karim, M.T. Hassan, M. Jeon, Terms-based discriminative information space for robust text classification, Information Sciences, Vol. 372, pp. 518-538, 2016

[5] G. Feng, J. Guo, Bing-Yi Jing, Tieli Sun, Feature subset selection using naive Bayes for text classification, Pattern Recognition Letters, Vol. 65, pp. 109-115, 2015

[6] M. Mohammadi, H. Parvin, E. Faraji, S. Parvin, Combining Classifier Guided by Semi-Supervision, Journal of Advances in Computer Research, Sari Branch, Islamic Azad University, Sari, I.R.Iran, Vol 8, No. 1, pp. 27-50, 2017.

[7] N. Ebrahimpour, F.S. Gharehchopogh; Z.A. Khalifehlou, New Approach with Hybrid of Artificial Neural Network and Ant Colony Optimization in Software Cost Estimation, Journal of Advances in Computer Research, Sari Branch, Islamic Azad University, Sari, I.R.Iran,Vol. 7, No. 4, pp. 1-12, 2016.

[8] M. Tavana, M. Mohammadi, H. Parvin, A Semi-Supervised Human Action Learning, Journal of Advances in Computer Research, Sari Branch, Islamic Azad University, Sari, I.R.Iran, Vol 7, No. 3, pp. 15-32, 2016.

[9] C. Shang, M. Li, S. Feng, Q. Jiang, J. Fan, Feature selection via maximizing global information gain for text classification, Knowledge-Based Systems, Vol. 54, pp. 298-309, 2013.

[10] Y. Lu, M. Liang, Z. Ye, L. Cao, Improved particle swarm optimization algorithm and its application in text feature selection, Applied Soft Computing, Vol. 35, pp. 629-636, 2015.

[11] R. Hu, B.M. Namee, S.J. Delany, Active learning for text classification with reusability, Expert Systems with Applications, Vol. 45, pp. 438-449, 2016.

[12] A. Askarzadeh, A Novel Metaheuristic Method for Solving Constrained Engineering Optimization Problems: Crow Search Algorithm, Computers and Structures, Vol. 169, pp. 1-12, 2016.

[13] Martin, Instance-Based Learning: Nearest Neighbour with Generalisation, Doctoral dissertation, University of Waikato, 1995

[14] M. Hasanlou, F.S. Gharehchopogh, Software Cost Estimation by a New Hybrid Model of Particle Swarm Optimization and K-Nearest Neighbor Algorithm, Journal of Electrical and Computer Engineering Innovations, pp: 49-55, Vol: 4, No: 1, 2016.

[15] E.E.Miandoab, F.S. Gharehchopogh, A Novel Hybrid Algorithm for Software Cost Estimation Based on Cuckoo Optimization and K-Nearest Neighbors Algorithms ", Engineering, Technology & Applied Science Research, Vol: 6, No: 3, pp. 1018-1022, 2016.

[16] F.S. Gharehchopogh, S.R. Khaze, I. Makeli, A New Approach in Bloggers Classification with Hybrid of K-Nearest Neighbor and Artificial Neural Network Algorithms ", Indian Journal of Science and Technology, Vol: 8, No: 3, pp: 237–246, Feb 2015.

[17] F.S. Gharehchopogh, Z.A. Khalifelu, Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing, 5th International Conference on Application of Information and Communication Technologies (AICT2011), IEEE press, pp. 1-4, Baku, Azerbaijan, 12-14 October 2011.

[18] A.K. Uysal, An improved global feature selection scheme for text classification, Expert Systems with Applications, Vol. 43, pp. 82-92, 2016.

[19] W. Zhang, X. Tang, T. Yoshida, TESC: An approach to TExt classification using Semi-supervised Clustering, Knowledge-Based Systems, Vol. 75, pp. 152-160, 2015.

[20] B.S. Harish, D.S. Guru, and S. Manjunath, Classification of Text Documents Using B-Tree, CCSIT 2012, Part II, LNICST 85, pp. 627-636, 2012

[21] K. Chen, Z. Zhang, J. Long, H. Zhang, Turning from TF-IDF to TF-IGM for term weighting in text classification, Expert Systems with Applications, Vol. 66, pp. 245-260, 2016.

[22] Y. Ko, J. Seo, Text classification from unlabeled documents with bootstrapping and feature projection techniques, Information Processing & Management, Vol. 45, Issue 1, pp. 70-83, 2009

[23] N. Shafiabady, L.H. Lee, R. Rajkumar, V.P. Kallimani, Nik Ahmad Akram, Dino Isa, Using unsupervised clustering approach to train the Support Vector Machine for text classification, Neurocomputing, Vol. 211, pp. 4-10, 2016.

[24] D. Wang, H. Zhang, Rui Liu, Weifeng Lv, Datao Wang, t-Test feature selection approach based on term frequency for text categorization, Pattern Recognition Letters, Vol. 45, pp. 1-10, 2014

[25] H. Uguz, A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, Knowledge-Based Systems, Vol. 24, Issue 7, pp. 1024-1032, 2011.

[26] H. Majidpour, F.S.Gharehchopogh, An Improved Flower Pollination Algorithm with AdaBoost Algorithm for Feature Selection in Text Documents Classification, Journal of Advances in Computer Research, Articles in Press, Accepted Manuscript, Available Online from 03 April 2017.

[27] Reuters Data Set, http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection, [Last Available :2016.04.05].

[28] Data Sets for Single Label Text Categorization, http://ana.cachopo.org/datasets-for-single-label-text-categorization, [Last Available :2016.04.05].

[29] K.K. Bharti, P.K. Singh, Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering, Expert Systems with Applications, Vol. 42, Issue 6, pp. 3105-3114, 2015.

[30] R.S. Michalski, I. Bratko, M. Kubat, Machine Learning and Data Mining: Methods and Applications, New York: Wiley, 1998.