Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 8, No. 4, November 2017), Pages: 73-86
www.jacr.iausari.ac.ir

A New Hybrid Model of K-Means and Naïve Bayes Algorithms for Feature Selection in Text Documents Categorization

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

Received: 2016/09/18; Accepted: 2016/11/25

Abstract

With increasing speed of information and documents on the Web, our need to classify them in different categories and clusters is more necessary. Clustering tries to find related structures in data sets which they are not categorized, yet. Concerning the needs, a new approach for text documents categorization is presented in this paper which includes three phases: pre-processing documents and selection feature, K-Means clustering and Naïve Bayes (NB) optimization. The proposed model uses K-Means and NB algorithms that utilize K-Means algorithm to find minimum distances between features from the center of clusters and NB algorithm for computing the probability of each feature into documents and using them to cluster features, separately. The proposed model optimizes performance of K-Means algorithm by using NB properties in clustering. Therefore, the model overcomes to the challenges of labeling different documents and origin of K-Means algorithm which it refers to categorizing text documents as un-supervised model. Finally, the experiment results of proposed model and K-Means algorithms are evaluated based on evaluation methods and are compared in validated datasets.

Keywords: *Text Categorization, Machine Learning, Feature Selection, K-Means Algorithm, Naïve Bayes Algorithm*

1. Introduction

Clustering is one of the text categorization methods as an un-supervised learning method. Clustering is one of the most applicable domains of data mining which aims to classify entities, items or features [1,2]. Features of each cluster are similar together and features of two different clusters are unlike each other [3]. This research considers clustering techniques to process natural language (English) called clustering and categorizing features with our proposed model using hybrid of K-Means and NB algorithms for clustering optimization. NB algorithm is studied extensively by some researchers in text classification task [4]. Our proposed model uses K-Means algorithm to find the similarity between clustering features and NB algorithms calculate the probability of each feature to optimize text document clustering and optimization of centers of clusters. The rest of the paper is structured as follows: The next section is explains related works of text categorization. After that, K-Means NB algorithms are introduced. The proposed model and experiment results evaluation and model validation are presented, respectively. Finally, we conclude our paper and future works.

2. Related Works

In the recent decades, text categorization is an increasingly growing area; because databases and information contained in web pages are growing larger. Many text documents such as news and science databases etc., exist in the World Wide Web environments. Performing a manual categorization is a time consuming task which makes it a motivation in terms of researches. Researchers presented more attentions to categorize text documents, automatically. For example, [5] presented multi label NB approach to text classification which used principal component analysis method to extracts unrelated features and then preprocesses documents. In order to increase performance, they introduced GA to select subset of features which have the most similarity. They show that their proposed model has an acceptable performance. The subset feature selection is a well-known method to improve performances of text categorization [6]. Therefore, [7] providing a model based on NB algorithm which used the selected subset features increases high performance in the classification of text documents. Special feature of this model is used feature indexing by global selection indexing method.

In another related work, [8] utilizes GA and DE algorithms to text clustering. They use crossover and mutation operators to optimize the number of clusters. These experimental results show that GA is more effective than DE method. It also outperforms hybrid algorithm and considered more effective than both of them. Clustering is one of the most data mining techniques that extract rules from text documents. It's been realized that by using KNN to extract data similarity in the clusters that were presented in [9], was more effective than K-Means to cluster text documents. In their study, the aim of clustering optimization is minimizing or maximizing the number of clusters. Text categorization by means of K-Means algorithm is presented in [10]. The main goal of using K-Means algorithm is selecting the nearest documents to the center of the best cluster. Their experimental results show that K-Means algorithm has high performance in text clustering with the lowest runtime.

For high performance of K-Means algorithm, GA is used to select related features to similar clustering [11]. They used GA to improve accuracy of clustering by creating and repeating different generations and decreasing computing operations. They also, show that the hybrid model has an improved performance of the K-Means algorithm. Other clustering algorithms are presented in [11]. The applicable areas of clustering are text mining, finding the similarity of documents, disambiguation and information retrieval, modeling language and text categorization.

A feature selection in text clustering is performed by binary particle swarm optimization (BPSO) [12]. Their proposed model used fitness based on adaptive inertia weight that is integrated with BPSO to control the exploration and exploitation of the particle in the search space. Opposition and mutation are integrated with BPSO which improve search capability and performance of clustering algorithm. Our research uses pre-processing phase to remove noisy features such as duplicated words, question remarks and etc. to improve the performance of feature selection and clustering within the aim of optimization of the center of clusters.

3. Text Categorization

Due to the explosive growth of electronic texts and documents on Internet in recent decades, text documents management is of particular importance. Text documents categorization is a suitable method for fast required data retrieval with a maximum speed among millions of documents. Text categorization or document categorization are used in Natural Language Processing (NLP) tasks such as filtering e-mails, specifying user desires, information extraction and etc., some of the texts categorization methods use keywords to classify texts in separate classes [13,14]. Other methods utilize the similarity of words (features) to clustering. Semantic relation is also another NLP technique for categorization based on the similarity among semantic relations. In the following section, two methods K-Means and NB algorithms are introduced.

3.1 K-Means Algorithm

One of the best well-known methods for clustering is K-Means algorithm [15, 16]. Despite the simplicity, K-Means is the basis for many of clustering methods. For this algorithm, many forms of expression is presented. But all of them have a routine to find a constant number of clusters and try to estimate and specify centroids as centers of clusters. The centroid points are known as average points and specifying sample data to each cluster with minimum distance of centroids. Fitness function of K-Means algorithm is formulated as Equation (1):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

Where Cj denotes the centroids of clusters and Xi (x1, x2, x3,…, xn) are vector of data samples (features), K is the number of clusters and n is number of features. K-Means algorithm firstly selects K member (number of clusters) randomly from among N-members and considers them as centers of clusters (centroids). Then, N-K remaining members are assigned to the nearest cluster. After the allocation of all members, centroids of cluster centers are calculated again and members according to their similarity amount (distance) are allocated to one of the clusters and this will continue until the cluster centers are constant. By repeating this procedure many times, new centroids are calculated by minimizing the data and accordingly features to new clusters are allocated. K-Means algorithm is one of the common methods of clustering [17]. Despite many advantages such as high speed and the ease of implementation, this method cannot reach an optimal solution for the problem proposed.

3.2. Naïve Bayes Algorithm

One of the most important tools to implement data mining techniques is NB algorithm [7]. NB algorithm explicitly on the probability of various hypotheses, such as NB classifier including efficient algorithms which is possible for some learning problems [5]. In this method, in order to maintain extracted information of learning it is required to calculate and store probability sets of documents and subjects and use to text categorization.

The NB reasoning method is a conditional probability technique. It assumes that features are categorized and represented by a vector X=(x1, …, xn) . Where n represents the number of features (independent variables), and the probabilities are assigned to this instance as follows:

$$p(C_k \mid x_1, ..., x_n) \tag{2}$$

Equation (2) implies that for each K there are possible clusters or classes that exist. The NB classifier is the function that assigns a class label $\hat{y}=C_k$ for K clusters as follows:

$$\overline{y} = \arg\max_{k \in \{1,...,k\}} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{3}$$

This method is appropriate to classify the problems in the real world. For instance, this method is implemented in medicine [18] and text categorization [19] and is considered as effective as other methods such as decision trees and artificial neural network [20]. We use this method to optimize centroids of clusters.

4. Proposed Model

K-Means clustering algorithm is an un-supervised method for categorization and clustering text documents. The features of this algorithm lay behind the fact that there is no need to label texts and assign un-labeled texts to categories or clusters. The objective function regarding this algorithm is finding minimum distance between features (words) which are centers of clusters (centroids). In contrast, NB algorithm is a supervised learning method that is used to classify text documents using various labeled texts. The objective function of this method is calculating the probabilities of features in each document and utilizing the value of these probabilities in text categorization.

Here, we face the challenge of proposing a hybrid model of two algorithms to categorize text documents. The first challenge is the process of creating a model to categorize and cluster by using an un-supervised model. The second challenge is creating a hybrid model without the need for labeling text documents. Actually, we try to cluster documents by K-Means algorithm and improve the quality of clustering and increase accuracy by calculating probabilities of features. To solve these issues, the proposed model is presented for finding the minimum distances between features by using K-Means algorithm and BN algorithm by computing the probability of each feature in documents and then optimize clustering based on the features' probabilities. Since the proposed model is an un-supervised learning model, there is no need to label documents.

The optimal model to address these issues is proposed in text documents categorization and accordingly improves the performance of the feature selection by using composition of K-Means and NB algorithms in three phases: pre-processing and feature selection, K-Means clustering and NB in clustering optimization. The proposed model is achieved after pre-processing the data and removing noisy data e.g. duplicate and waste words try to cluster related documents together. In the next stage of the learning process, the system increases the K-Means algorithm clustering accuracy. As mentioned above the clustering process will be optimized based on the probabilities of the features by implementing the NB algorithm. Furthermore, the results of the new approach applied in text documents categorization using validated datasets for text categorization [21] are compared. Clustering and optimizing phases continue to achieve the lowest error and fixed centroids for all clusters. The flowchart of our proposed model is shown in Figure 1.

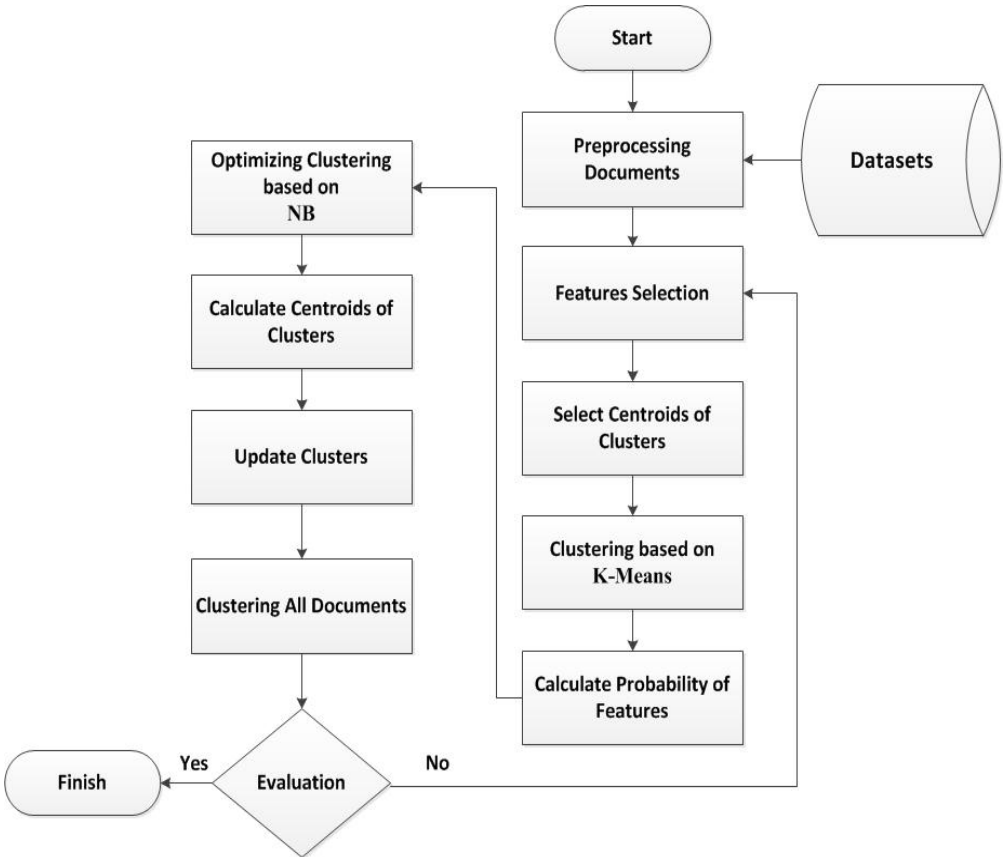

Figure 1. Flowchart of Proposed Model

As shown in Figure 1, we pre-process documents before clustering by removing un-useful words. Pre-processing phase involves removing redundant and duplicate words, conjunctions, news and question marks. The purpose of pre-processing is the data analysis and the remove of ineffective data. Pre-processing includes the correction of data in terms of disorder and noise and its goal is to obtain accurate classification.

The feature selection phase is performed after that pruning of documents by pre-processing procedure. In the second phase, K-Means algorithm calculates clusters, centroids of clusters and minimum distances to allocate features to the nearest K clusters. In order to increase performance of K-Means algorithm, NB is used to select related words with the closest probabilities in the proposed model. In the third phase, optimization by NB algorithm is generated and the features which are most likely (closest probability) are allocated in a specified cluster according to the probabilities. The centroids of clusters have to be updated and then clusters based on new centroids of clusters are optimized. This procedure is continued until all of documents are clustered. Evaluation and validation of the model is the final phase of our proposed model which is presented concerning evaluation measures. The quasi code of our proposed model is shown in Table 1.

Table1. Quasi code of Proposed Model

- Start
- Initialize Parameters (K, N, xi)
- Iteration for each Document
- Pre-processing Documents (j)
- Do (xi)
- -Select centroids Cj
- -Calculate
- -K-Means clustering
- -Calculate measures (Purity, Entropy)
- -NB: Calculate probability of xi: (P(Ck|xi), ŷ)
- -Update Centroids (Cj) by probability xi
- -Update clustering by probability xi
- -Optimize K-Means algorithm
- -Calculate measures (Purity, Entropy, Accuracy)
- While constant centroids (Cj)
- Evaluation (true)
- Finish

In the clustering level of K-Means algorithm, the features are constructed in a way that K is the number of clusters and the random centroids are generated as for each cluster. After initializing all clusters, the nearest centroids of all features are calculated and updated for each feature, again. Optimization is based on NB calculated probabilities to update means and clusters. This model is implemented in VC#.Net 2015 programming language, datasets [22] and parameters are presented in Table 2.

Table2. Datasets Description and Parameters

Datasets	Number of Documents	K-Values
Reuters-21578 R8	5485	3,4,5,6,7
Reuters-21578 R52	6532	3,4,5,6,7
Cade 12	27322	3,4,5,6,7
WebKB	2803	3,4,5,6,7
Reuters-21578	100,250,500	3,4,5,6,7

5. Evaluation Methods and Results

In order to evaluate and validate the proposed model, performance measures (Entropy, Purity and Accuracy) are introduced. The proposed model along with K-Means algorithm are calculated regarding the performance measures and evaluated according to various datasets selected to conduct this research. Purity in equation (4) and Entropy in equation (5) are formulated as shown below.

$$Purity = \sum_{j=1} \frac{n_j}{n} \arg\max P_{ij} \quad (4)$$

$$Entropy = -\frac{1}{\log c}\sum_{j=1}^{m}\frac{n_j}{n}\sum_{i=1}^{m}P_{ij}\log P_{ij} \quad (5)$$

Where P_{ij} is the probability of a document from cluster j which is belonging to class *i*; m is number of clusters; n_j is number of documents in cluster j and n is number of all of documents. Accuracy measure is formulated as Equation (6).

$$Accuracy(A) = \frac{(TP + TN)}{Total\_Sample} \quad (6)$$

Where a true positive (TP) decision assigns two similar documents to the same cluster, a true negative (TN) decision assigns two dissimilar documents to different clusters and Total samples refers to all features available for clustering. Table 3 presents the results of proposed model in Reuters-21578 R8 dataset. 5485 text documents exist in this dataset. The results are based on Entropy and Purity of clustering used to measure goodness of a clustering and K is number of clusters. Entropy is an internal measure for the quality of a clustering. The aim is to obtain lowest value of Entropy and highest value of Purity in all experiment results.

Table 3. Results of Our Approach on Reuters-21578 R8 Dataset

Datasets	Number of Documents	K-Values	Purity %		Entropy %	
			Proposed Model	K-Means	Proposed Model	K-Means
Reuters-21578 R8	5485	3	91.60	86.80	72.70	77.00
		4	90.10	85.40	71.70	75.70
		5	83.40	79.20	67.30	70.00
		6	77.70	73.90	63.50	65.00
		7	71.60	68.20	59.40	59.70

As shown in Table 3 with increasing K for clustering, more different documents can participate in decision-making of clustering, accordingly performance of results is somehow reduced, since more documents by lower similarity score are involved in clustering that force to reduce the efficiency of clustering documents. As shown, the proposed model is able to increase the performance of our model rather than K-Means algorithm due to the lower Entropy and higher Purity values obtained. Table 4 shows the results of performance measures in Reuters-21578 R52 dataset. This dataset include 6532 text documents and K is number of clusters.

Table 4. Results of Our Approach on Reuters-21578 R52 Dataset

Datasets	Number of Documents	K-Values	Purity %		Entropy %	
			Proposed Model	K-Means	Proposed Model	K-Means
Reuters-21578 R52	6532	3	88.50	83.90	70.60	74.40
		4	84.80	80.40	68.20	71.10
		5	74.30	70.70	61.20	62.10
		6	72.80	69.30	60.20	60.8
		7	70.50	67.10	58.60	58.70

As shown in Table 4, Entropy value is lower than others so the result of our model is better and also Purity value is high which present the performance of our model. By the way, Cade 12 dataset include 27322 text documents. Table 5 shows the results of algorithms and our proposed model for dataset Cade 12.

Table 5. Results of Our Approach on Cade 12 Dataset

Datasets	Number of Documents	K-Values	Purity %		Entropy %	
			Proposed Model	K-Means	Proposed Model	K-Means
Cade 12	27322	3	88.10	83.5	70.40	74.00
		4	85.80	81.40	68.80	72.00
		5	70.30	67.00	58.50	58.60
		6	67.00	63.90	56.30	55.70
		7	55.50	53.10	48.60	45.70

As shown in Table 5, Entropy and Purity values show excellence of our model for example for k=5 entropy is % 70.3 and Entropy is %58.5. The results of proposed model and K-Means algorithm for dataset WebKB are computed in Table 6. This dataset is involved 2803 text documents.

Table 6. Results of Our Approach on WebKB Dataset

Datasets	Number of Documents	K-Values	Purity %		Entropy %	
			Proposed Model	K-Means	Proposed Model	K-Means
WebKB	2803	3	94.80	89.80	74.90	79.80
		4	93.20	88.30	73.80	78.50
		5	78.20	74.30	63.80	65.40
		6	69.90	66.60	58.30	58.30
		7	62.30	59.40	53.20	51.60

The result of our proposed model in Table 6 outperforms K-Means algorithm based on Purity and Entropy measures. For example, purity (%94.8), entropy (%74.9) are better results of K-Means algorithm (purity=%89.8 and entropy %79.8) and K=3. By increasing the number of clusters, the results of measures show that our proposed model is more effective than K-Means algorithm and optimization of K-Means algorithm with NB (our proposed model) is reasonable. Based on experiment results, our proposed model on dataset Reuters-21578 R8 is more effective than the other conducted datasets. This may be due to the content of clustering documents and features available in it. Also, this indicates that convergence of features in the dataset (Reuters-21578 R8) which is higher than the other used datasets which generate these results.

As shown in Figure 2, the results of proposed model and K-Means algorithm are compared. Performance evaluation measure is based on Entropy value. The lower value of Entropy shows more efficiency of the model and determines uncertainty for a cluster set. As shown, our proposed model outperforms K-Means algorithm; because our proposed model in all datasets has lower entropy than K-Means algorithm. Number of clusters is K=3, 4,5,6,7.

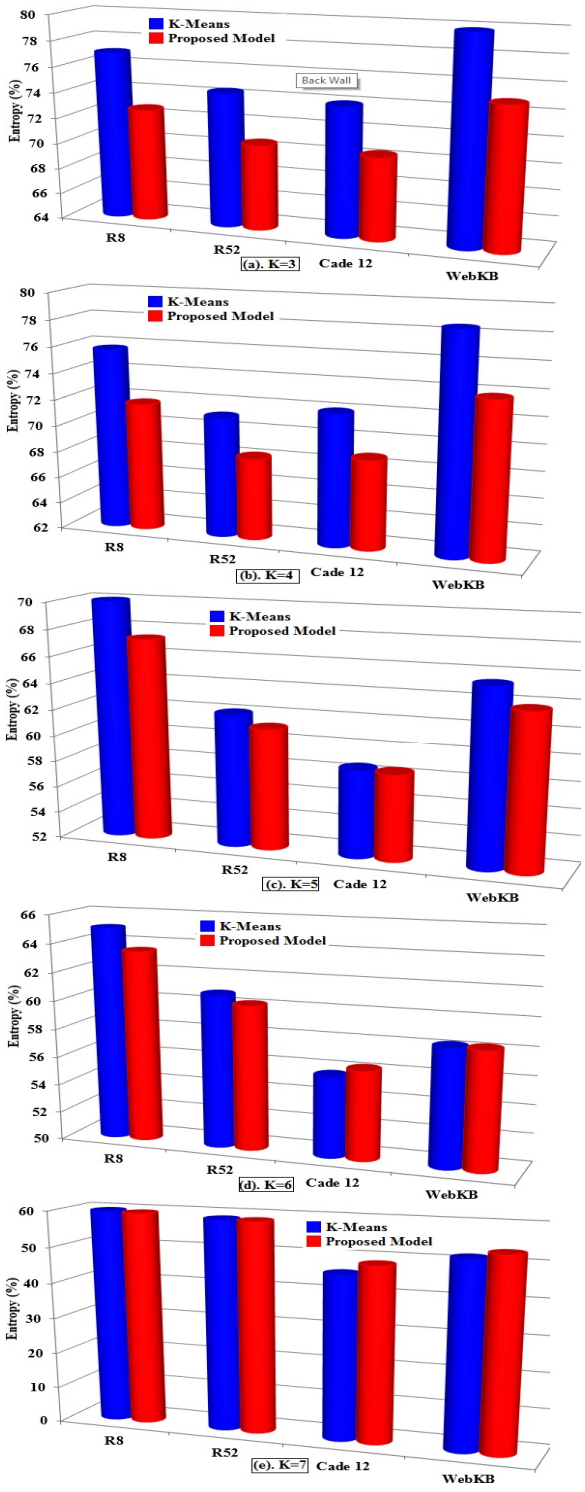

Figure 2. Comparison of Entropy Values on Different Datasets

The entropy identifies the irregularities a set of data. As it can be seen in Figure 2, the proposed model has less irregularity at the data categorization in comparison with the K-Means model and as a result, it is more accurate in the classification of the documents. In Figure 3, the results of our proposed model and K-Means algorithm are compared. Performance evaluation measure is based on Purity measure in different

datasets. As shown, proposed model outperforms K-Means algorithm. Number of clusters is K=3, 4,5,6,7.

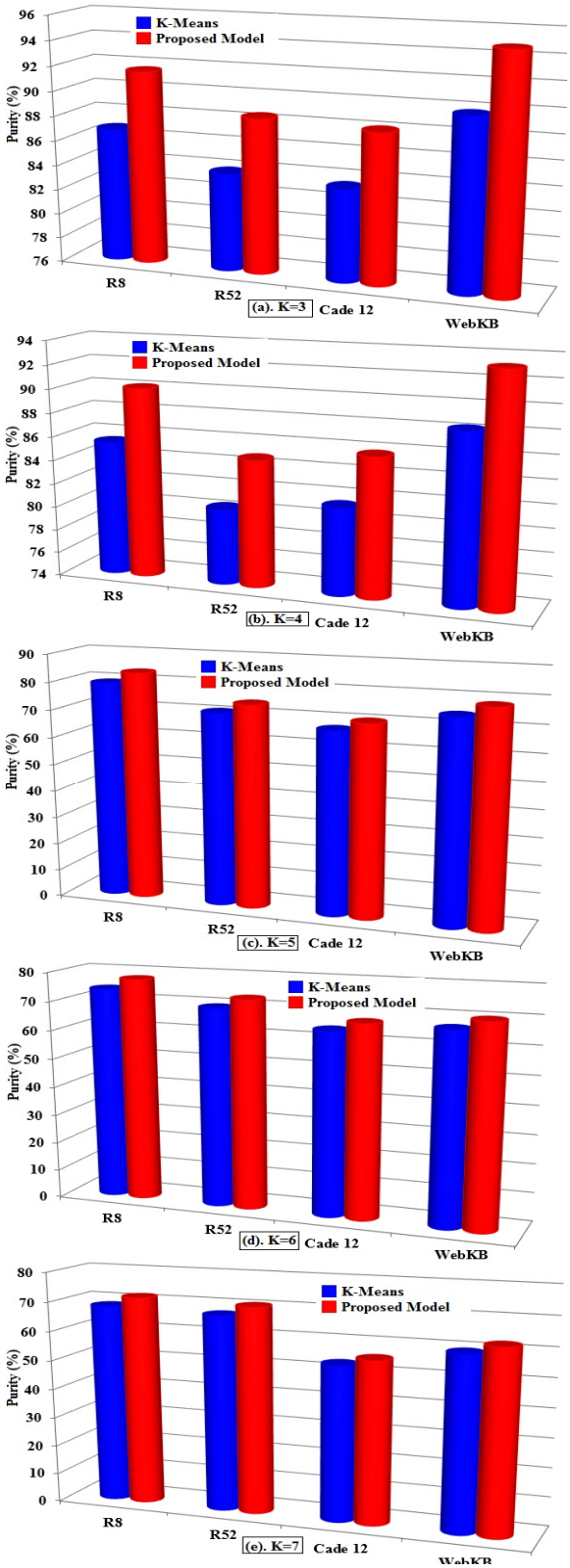

Figure 3. Comparison of Purity values on Different Datasets

As shown in Figure 3, the proposed model presents more excellence than K-Means algorithm; because in all datasets this model has higher Purity than K-Means algorithm. In Table 7, the results of our proposed model are compared with other systems in the same datasets with similar K-values.

Table7. Results Comparison with other Approaches in Reuters-21578 Dataset

System	K-Values	Purity %	Entropy %
K-Means [21]	3	76.85	76.40
	4	68.23	69.65
	5	55.92	60.74
Habibpour and Khalilpour [21]	3	87.32	61.12
	4	72.37	56.85
	5	67.15	49.17
Proposed Model	3	93.30	60.80
	4	83.20	53.10
	5	78.80	50.20

The model has been tested over 4 databases corresponding to commonly used benchmarks. Better quality of clustering shows higher Purity and lower Entropy. Also, if clustering results match exactly with existing clusters, then Entropy value to be 0 and Purity to be 1. Results in Table 7 shows that our proposed model outperforms the model [21] and K-Means algorithm. One of the problems in text documents clustering is selecting similar features between different documents. Often, because of irrelevant and redundant features, clustering efficiency will be reduced. Thus, pre-processing of documents is essential. That's why our proposed algorithm improves the clustering results and increases performance of the proposed model. The next reason for improved results and high performance of our approach compared to other clustering algorithms is using probability values to assign features to their own nearest clusters and optimizing centroids of clusters updated by NB algorithm in each run of our proposed algorithms. The end condition of the proposed model is constant centroids of clusters after optimization. Also, Table 8 is a comparison between our proposed model and other approaches to evaluate accuracy of systems, significantly. The aim is to achieve the highest Accuracy value to show its high performance based on an Accuracy measure.

Table 8. Result Comparison based on Accuracy Value

Datasets	System	Accuracy %
Reuters-21578 R8	GA [23]	88.63
	K-Means	71.65
	Proposed Model	90.23
WebKB	GA [23]	87.49
	K-Means	73.32
	Proposed Model	85.75

In order to demonstrate the accuracy rate, the proposed model has been compared with GA model [18]. GA model [18] is a good choice for comparison with the proposed model due to the dataset of Reuters21578 and WebKB. As shown in Table 8, our proposed model results are more promising than the proposed model in [21] and K-Means algorithm over Reuters-21578 R8 dataset. This particular result is consistent with a previous study where it was also found that the K-Means text categorization performed the worst in [21] and [23]. Regarding the inverse document frequency, it is concerned essential because it reduces the importance of common words existing in a lot of documents. Yet, our experiments show that their categorization improved the performance, slightly.

However, manual clustering is a time consuming job and needs huge human effort which makes it un-scalable with respect to the high growing speed of the documents in databases, web and other text warehouses. Therefore, there is a need for automated clustering systems. This motivation made us design and present a new approach in automated clustering by optimization previous methods e.g. K-Means optimization in clustering. Key strengths of our proposed model is using NB properties in hybrid text categorization model which optimize performance of this proposed model by calculating probabilities of features after pre-processing documents. Another advantage is the pre-processing phase which improves quality of clustering by removing noisy features. The noisy features are redundant and duplicate words, conjunctions, news and question marks and etc. With K-Means classifier, we performed experiments to select the best number of k means and the best feature space transformation.

K-Means is one of the simplest base line algorithms which uses un-supervised learning approach to solve clustering challenges. The advantage is that it works really well with large datasets. As a disadvantage, it has strong sensitivity to outliers and noise features so that pre-processing phase is necessary to overcome limitation and to achieve higher performance. It has a low capability to pass the local optimum and low performance of this algorithm. Using NB properties in clustering improvement solve this issue as one of advantages of our optimization. Another advantage of using NB algorithm is based on conditional probabilities and requires a small amount of data to estimate the parameters necessary for clustering. Other clustering algorithms with better features tend to be more expensive. In this case, K-Means algorithm becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where using NB algorithm can be applied to clustering optimization. Other advantage of this model is no needs to labeling documents and training phase due to our designated un-supervised learning model.

Un-supervised NB for data clustering with mixture of K-Means algorithm frequently known as clustering is a descriptive task with many applications e.g. pattern recognition and information retrieval that can be also used as a pre-processing task in the so-called knowledge discovery from data process. The algorithm that we use in this paper can still be improved. For instance, during the model construction we use an un-supervised approach for estimating the parameters and validating the models obtained. We consider that the results obtained, in compares with the K-Means algorithm and GA [23], are promising.

6. Conclusion and Future Works

Nowadays, with ever increasing information speed on Internet more than past time the need to categorize texts documents is considered more necessary. Text categorization accelerates fast search and documents retrieval from a large number of information. In fact, categorization is assigning documents to related groups or clusters. Concerning the issue of text clustering, a new approach to document clustering is presented. In this paper, a method to optimizing performance of K-Means clustering algorithm that involves three-phase clustering is presented. In the first phase, documents are pre-processed and features are selected. In the second phase, clustering based on K-Means algorithm by specifying several centroids and then allocating features to each cluster is carried out. In the third phase, improvement of clustering is based on the probability values of features using NB algorithm optimized. Using NB properties in clustering improvement solve this issue as one of advantages of our optimization. Another advantage of using NB algorithm is based on conditional probabilities and requires a small amount of data to estimate the parameters necessary for clustering. The evaluation measures of clustering quality are based on Purity, Entropy and Accuracy which indicates that our proposed model in compares with the K-Means algorithm and GA are promising. As future works, optimization of clustering is suggested based on natural language processing techniques such as semantic and syntactic relations and using other meta-heuristic methods to improve automated clustering systems.

References

[1] 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.

[2] 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, Autumn 2015.

[3] F.S. Gharehchopogh, Z.A. Khalifelu, "Study on Information Extraction Methods in Unstructured Data: Text Mining versus Natural Language Processing", AWERProcedia Information Technology & Computer Science Journal, Vol:1, pp.1321-1327, 2012.

[4] J. Chen, H. Huang, S. Tian, and Y. Qu, (2009), Feature Selection for Text Classification with Naïve Bayes, Expert Systems with Applications, vol. 36, no. 3, pp. 5432-5435.

[5] M. L. Zhang, J. M. Pena, V. Robles, (2009), Feature Selection for Multi-Label Naïve Bayes Classification, Information Sciences, Vol. 179, Issue. 19, pp. 3218-3229.

[6] N. M. Darani, V.Ahmadi, Z. S. Eskandari, M. Yousefikhoshbakht, (2013), Solving the Capacitated Clustering Problem by a Combined Meta-Heuristic Algorithm, Journal of Advances in Computer Research, Vol 4, Issue: 1, pp. 80-100, Winter 2013.

[7] G. Feng, J. Guo, B. Y. Jing, T. Sun, (2015), Feature Subset Selection Using Naïve Bayes for Text Classification, Pattern Recognition Letters, Vol. 65, pp. 109-115.

[8] Y.K. Meena, Shashank, V.P. Singh, (2012), Text Documents Clustering using Genetic Algorithm and Discrete Differential Evolution, International Journal of Computer Applications, Vol. 43, No. 1, pp. 16-19.

[9] C. Luo, Y. Li, S.M. Chung, (2009), Text Document Clustering Based on Neighbors, Data & Knowledge Engineering, Vol. 68, pp.1271–1288.

[10] Chen Y., Qin B., Liu T., Liu Y., Li Sheng, (2010), The Comparison of SOM and K-Means for Text Clustering, Computer and Information Science 3(2), 268–274.

[11] Verma H., Kandpal E., Pandey B., Dhar J., (2010), A Novel Document Clustering Algorithm Using Squared Distance Optimization Through Genetic Algorithms", International Journal on Computer Science and Engineering, Vol. 02, No. 05, pp. 1875-1879.

[12] K. K. Bharti, P. K. singh, (2016), Opposition Chaotic Fitness Mutation Based Adaptive Inertia Weight BPSO for Feature Selection in Text Clustering, Applied Soft Computing, Vol. 43, pp. 20-34.

[13] 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, Winter 2015.

[14] 14.M.h Haghiri, H. Hassanpour, (2011), Using Supervised Clustering Technique to Classify Received Messages in 137 Call Center of Tehran City Council, Journal of Advances in Computer Research, Vol 2, Issue: 3, pp. 15-24, Summer 2011.

[15] B. MacQueen,(1967), Some Methods for Classification and Analysis of Multivariate Observations, 5th Berkley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281-297.

[16] F.S. Gharehchopogh, N. Jabbari, Z. Ghaffari_Azar, "Evaluation of Fuzzy K-Means and K-Means Clustering Algorithms in Intrusion Detection Systems ", International Journal of Scientific & Technology Research (IJSTR), Vol: 1, Issue 11, pp.66-72, December 2012.

[17] M. Yao, D. Pi, X. Cong, (2012), Chinese Text Clustering Algorithm based K-Means, International conference on Medical Physics and Biomedical Engineering, Physics Procedia, Vol. 33, pp. 301-307.

[18] P.L. Geenen, L.C. vander Gaag, W.L.A. Loeffen, A.R.W. Elbers, (2011), Constructing naive Bayesian classifiers for veterinary medicine: A case study in the clinical diagnosis of classical swine fever, Research in Veterinary Science, Vol. 91, Issue. 1, pp. 64-70.

[19] J. Wu, S. Pan, Z. Cai, P. Zhang, C. Zhang, (2015), Self-adaptive attribute Weighting for Naïve Bayes Classification, Expert Systems with Applications, Vol. 42, Issues. 3, pp. 1487-1502.

[20] N. Ebrahimpour, F.S. Gharehchopogh; Z. A.Khalifehlou, (2016), New Approach with Hybrid of Artificial Neural Network and Ant Colony Optimization in Software Cost Estimation, Journal of Advances in Computer Research, Vol 7, Issue: 4, pp. 1-12, Autumn 2016.

[21] R. Habibpour, k. Khalilpour, (2014), A New Hybrid K-Means and K-Nearest Neighbor Algorithms for Text Document Clustering, International Journal of Academic Research Part A; 6(3), 79-84.

[22] Reuters Datasets, (2016), Available at: http://ana.cachopo.org/datasets-for-single-label-text-categorization, [Last Available: 2016.2.25].

[23] H. J. Escaltane, M.A.G. Limon, A.M. Reyes, M. Graff, M. M. Gomez, E.F. Morales, and J. M. Carranza, (2015), Term-Weighting Learning via Genetic Programming for Text Classification, Knowledge-based Systems, Knowledge-Based Systems, 83:176-189.